

What do you base that conclusion on?

Grounding explainable AI in human dialogue strategies

Alexander Berman

Department of Philosophy, Linguistics and Theory of Science



UNIVERSITY OF GOTHENBURG

2026

Copyright 2026 by Alexander Berman. The author is supported by grant 2014-39 from the Swedish Research Council for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

Cover design by the author
Printed in Sweden by Stema Specialtryck AB
Publisher: University of Gothenburg (Dissertations)

Typeset in Baskervald X

ISBN 978-91-8115-771-0 (print)
ISBN 978-91-8115-772-7 (PDF)

Distribution:
Department of Philosophy, Linguistics and Theory of Science
Box 100, SE-405 30 Gothenburg

In memory of my dear friends Lotta & Björn

Abstract

When human decisions are assisted by predictions from artificial intelligence (AI) models, users' ability to understand the basis of AI outputs can be important for various reasons, such as assessing the reliability of specific predictions. One way to achieve such understanding is by letting AI systems explain their predictions. This thesis explores how AI systems can be designed to emulate how humans manage explanations in dialogue.

The first part of the thesis studies expert explanations for medical judgements in clinical settings without AI assistance as well as lay explanations in an experimental setting with AI assistance. By applying a dialogue distillation methodology, collected human-human interactions are rewritten into analogous human-AI dialogues, thereby revealing capabilities that a conversationally explainable AI system would need to possess to emulate human explanatory behaviours. Based on the finding that human interlocutors often explain judgements argumentatively (as claims supported by premises) and enthymematically (by omitting one or more premises), the thesis proposes a method for extracting arguments from generalised linear models, a popular type of predictive model, and demonstrates that the method can be used to generate enthymematic explanations that invite inferences that correctly reflect the actual reasoning of the model. A design workshop with orthopaedic surgeons is also performed, indicating that users find the generated explanations informative and that they can correctly interpret generated explanations.

In the final part of the thesis, observed human dialogue strategies are formally modelled in a novel framework for explanatory dialogue management based on information-state updates conceived as linear implications. The proposed framework accounts for many of the observed phenomena and strategies, including complex explanantia, chained inferences, forward expansions and signalling of presupposition violations and answer unavailability. Future research challenges associated with emulating observed human explanation strategies are also identified and discussed.

Sammanfattning

När mänskliga beslut assisteras av förutsägelser gjorda genom modeller baserade på artificiell intelligens (AI) kan det av flera skäl vara viktigt att användare har möjlighet att förstå vad sådana förutsägelser baseras på, bland annat för att kunna bedöma specifika förutsägelser tillförlitlighet. Ett sätt att uppnå sådan förståelse är att AI-system förklarar sina bedömningar. Denna avhandling utforskar hur AI-system kan utformas för att efterlikna hur människor hanterar förklaringar i dialog.

Avhandlingens första del studerar experters förklaringar till medicinska bedömningar i kliniska miljöer utan AI-assistans samt lekmanförklaringar i en experimentell miljö med AI-assistans. Genom tillämpning av en dialogdestilleringsmetodologi skrivs insamlade människa–människa-interaktioner om till motsvarande människa–AI-dialoger. Därigenom belyses de förmågor som ett konversationellt förklarbart AI-system behöver ha för att efterlikna mänskliga förklaringsbeteenden. Utifrån observationen att mänskliga konversationsdeltagare ofta förklarar bedömningar argumentativt (som påstående understödda av premisser) och entymematiskt (genom att utesluta en eller flera premisser) föreslår avhandlingen en metod för att extrahera argument från generaliserade linjära modeller, en populär typ av prediktiv modell, och visar att metoden kan användas för att framställa entymematiska förklaringar som korrekt återspeglar hur modellen faktiskt resonerar. En designworkshop med ortopedier genomförs också, vilken indikerar att användare uppfattar framställda förklaringar som informativa och att de kan tolka framställda förklaringar korrekt.

I avhandlingens avslutande del modelleras observerade mänskliga dialogstrategier formellt i ett nytt ramverk för hantering av förklaringsdialog där informationstillståndsuppdateringar behandlas som linjära implikationer. Det föreslagna ramverket hanterar många av de observerade fenomenen och strategierna, såsom komplexa explanantia, inferenskedjor, framåtriktade utvidgningar samt signalering av presuppositionsbrott och avsaknad av svar. Framtida forskningsutmaningar kopplade till att efterlikna observerade mänskliga förklaringsstrategier identifieras och diskuteras också.

Acknowledgements

When I was a child, my grandfather Zygmunt occasionally asked me about school. Which subjects did I enjoy? What triggered my curiosity? In retrospect, I wonder if my grandfather's focus on inner motivation and drive rather than external rewards or measurable results has influenced some of my life choices. Pursuing a doctoral degree is effortful and laborious, without evident payoffs. Yet, when I decided to apply for a doctoral position, the opportunity to spend four or five years digging my own rabbit hole and exploring potential tunnel systems connecting various research fields and problems seemed rewarding enough.

This book is a concrete outcome of that choice, narrated as if the path from research questions to answers (and follow-up questions) was clear and straightforward. In reality, however, the digging adventure followed a more meandering trail. I want to thank my supervisors Christine Howes and Jean-Philippe Bernardy for letting me explore territories freely while simultaneously helping me structure my tasks and activities, as well as for their support, advice and encouragement. I am also grateful for my opponent in the final seminar, Philipp Cimiano, who offered constructive and insightful guidance towards the end of the tunnel.

Among the people whom I have been fortunate to encounter in the subterranean tunnel systems and collaborate interdisciplinarily with during these years, I particularly want to thank Catharina Parai, Karl de Fine Licht and Vanja Carlsson, whose enthusiasm in identifying and jointly exploring overlapping research interests, as well as their dedicated commitment and generous attitude, has been immensely valuable for me.

Several additional individuals have offered valuable insights, suggestions and encouragement along the way; these include Kerstin Jacobsson, Ida Seing, Ylva Wallinder, Katarina Hollertz, Karolina Sztandar-Sztanderska, Marcin Zarod, Christian Munthe, Peter Ljunglöf, Linus Holmberg, Gerlof Bouma, Federico Cabitza, Liila Taruffi, Mattias Wahde, Freddy Lecue and Måns Molander. I also appreciate the stimulating discussions with people with different areas of expertise which I have encountered in the process. Thank you Dina Koutsikouri, Jonna Bornemark, Muhammad Suffian, Olle Häggström, Andreas Xydis, Hannah Ruschemeier, Jacob Livingston Slosser, Fredrik Johansson, Amelie Robrecht, Nils Johannesson, Adam Ek, Antonio Molin and Lena Enqvist for making my work so enjoyable.

Various colleagues at the Department of Philosophy, Linguistics and Theory of Science (FLoV) have been influential along this journey. Shalom Lappin warmly welcomed me to the Centre for Linguistic Theory and Studies in Probability

(CLASP) and immediately offered his guidance and support. Staffan Larsson lured me into dialogue systems research many years ago, which I am grateful for, as well as for our continued collaborations. Eleni Gregoromichelaki has acted as an informal mentor and endless knowledge resource in linguistic theory. I am also very thankful for the stimulating discussions with Ellen Breitholtz, Vlad Maraev, Amy Han Qui, Robin Cooper, Simon Dobnik, Erik Lagerstedt, Anna Folland, Mattias Granberg Olsson and Orvar Lorimer Olsson, as well as for the administrative and technical support from Susanna Myyry, Monica Havström, Sandra Schriefer, Cecilia Groglopo, Robert Adesam and Kaj Ailomaa. I also want to thank Erica Schweder from the university's legal and information management unit for valuable legal advice.

Several individuals have kindly contributed in various ways to the research presented in this thesis, including the colleagues who participated as subjects in pilot studies, as well as Erik, Filipa and Amy who volunteered in photo sessions for marketing my experiment.

My previous work in industry and the arts has offered inspiring and creative environments, and helped me prepare for future challenges. I particularly want to mention Henryk Rubinstein at Textalk who gave me the opportunity to explore my interests and learn new skills in a commercial setting when I was still very young. I also want to show my appreciation towards my colleagues at Volvo Technology, Talkamatic and in AI_am for the stimulating and rewarding collaborations.

Several organisations and individuals have invited me to present my research to general and non-academic audiences, which I have found stimulating and instructive. In this context, I particularly want to mention Petra Andreen, Ulrik Lork, Jan Andrée, Isabelle Kristensen, Alexandra Oljans Ahlin, and the staff at the International Science Festival Gothenburg.

My friends have played crucial roles in making this thesis happen, and, perhaps even more importantly, for filling these years with light, warmth and connection. Thank you Svante, Richard, Soheil, Filip, Malin, Jonas, Cia, David and all my other wonderful friends.

Finally, I want to mention my dear and beloved family. For as long as I can remember, my parents Denise and Leon have believed in my abilities and encouraged me to pursue my interests. My brother Robert is an endless source of stimulating conversation and warm encouragement. Lastly, two persons have endured with my occasional weariness, mood swings and perhaps somewhat nerdy and sometimes work-related choice of conversation topics (and will probably need to continue enduring despite the completion of this thesis). Stina and Artur: thank you for not asking why I spent so much time in the rabbit hole, and more generally for being such wonderful company.

Contents

1	Introduction	5
1.1	Explainable AI in the era of large language models	6
1.2	Adhering to human explanatory norms and strategies	6
1.3	Research questions	7
1.4	Methodologies and theoretical frameworks	8
1.5	Positioning the thesis in the research landscape	8
1.6	Outline of thesis	9
2	Philosophical and normative framework	11
2.1	What is an explanation?	11
2.2	Explanations in the context of AI	13
2.3	Normative aspects of explanatory communication	13
2.4	Arguments and other related concepts	15
2.5	Different types of AI explanations	16
2.6	Prerequisites for epistemically transparent AI	18
2.7	Choice of research direction	20
3	Explanations in non-AI-assisted medical dialogues	23
3.1	Related work	23
3.2	Data	24
3.3	Method	25
3.3.1	Technical assumptions	26
3.3.2	Normative assumptions	27
3.3.3	Iterative and incremental process	27
3.4	Results	28
3.4.1	Explanation triggers and query types	28
3.4.2	Types of explanantia	31
3.4.3	Explanatory underspecification	35
3.4.4	Turn-taking and complex explanantia	36
3.4.5	Chained inferences	37
3.4.6	Counter-argumentation	39
3.4.7	Hedging and epistemic stance	41
3.4.8	Answer unavailability	41
3.4.9	Ontic explananda	41
3.4.10	Feedback and grounding	42
3.4.11	Anaphora	43
3.4.12	Methodological elaborations	43

3.5	Discussion and conclusions	44
4	Lay explanations in dialogues involving AI-assisted personality assessments	47
4.1	Related work	48
4.2	Experimental setup	48
4.3	Recruitment of participants	53
4.4	Method of analysis	54
4.5	Results	54
4.5.1	Types of explananda	55
4.5.2	Explanation triggers and query types	57
4.5.3	Types of explanantia	58
4.5.4	Response strategies	63
4.5.5	Argumentative structure	64
4.5.6	Faithfulness and sources of evidence	66
4.5.7	Answer unavailability	67
4.5.8	Feedback and grounding	69
4.5.9	Anaphora	71
4.5.10	Turn-taking and complex explanantia	72
4.5.11	Ellipsis and explanandum co-referencing	73
4.5.12	Sub-topic structure	74
4.5.13	Reliability and epistemic stance	75
4.6	Limitations	77
4.7	Discussion and conclusions	77
4.8	Ethics declaration	80
5	Extracting arguments from interpretable models	81
5.1	Background and related work	82
5.1.1	Enthymemes	82
5.1.2	Explanation faithfulness	83
5.1.3	Effects of AI explanations on human understanding	84
5.1.4	Previous work on extracting arguments from ML models	85
5.2	Extracting arguments from GLMs	90
5.2.1	Model specification	90
5.2.2	Featurisation	91
5.2.3	Extracting data	91
5.2.4	Extracting warrants	93
5.3	Formal guarantees on correctness of invited inferences	94
5.3.1	Inferences from datum-based enthymeme	94
5.3.2	Inferences from warrant-based enthymeme	95

5.3.3	Assumptions concerning expected monotonicity	95
5.3.4	Relation to opaque models	96
5.4	Case study: Spinal surgery decision-support	96
5.4.1	Datasets	97
5.4.2	Feature encoding	98
5.4.3	Interface design	99
5.4.4	Choice of reference point	99
5.4.5	Generation of data and warrants	101
5.4.6	Evaluation	102
5.5	Limitations	108
5.6	Conclusions	108
5.7	Ethical considerations	110
6	Modelling explanatory dialogue management	111
6.1	Related work	112
6.1.1	Modelling and implementing explanatory and argumentative dialogue	112
6.1.2	Modelling dialogue management using information states	115
6.2	Preliminaries	116
6.2.1	Representing information about the predictive model	117
6.2.2	Representing the state of the dialogue	119
6.3	Dialogue modelling	121
6.3.1	Basic question answering and assertions	122
6.3.2	Basic explanations	125
6.3.3	Complex explanantia	129
6.3.4	Chained inferences	133
6.3.5	Forward expansions	135
6.3.6	Contradictory and neutral evidence	136
6.3.7	Framing of data	140
6.3.8	Confirmation questions	142
6.3.9	Presupposition violations	148
6.3.10	Answer unavailability	150
6.3.11	Negative understanding	151
6.4	Validation	152
6.5	Limitations	153
6.5.1	Providing source or type of evidence	153
6.5.2	Backing of warrants	154
6.5.3	Referring to related information	154
6.5.4	Addressing reference points in dialogue	154
6.5.5	Base-rate explanations	155

6.5.6	Generic warrant statements	156
6.5.7	Natural language generation	157
6.5.8	Natural language understanding	158
6.5.9	Rigidity	160
6.6	Conclusions and future work	160
7	Discussion and conclusions	163
7.1	Summary of results	163
7.2	Limitations	164
7.3	Future work	165
7.3.1	Addressing revealed dialogue modelling challenges	165
7.3.2	Analysing more data	166
7.3.3	Targeting broader range of ML models and approaches	166
7.3.4	Evaluating proposed dialogue model	167
7.4	Transparency in a broader perspective	167
7.5	The ultimate why-question	169
A	Original dialogue excerpts	171
B	Scenario specifications	173
C	BKOS: Formal specifications	175
C.1	Formalism for update rules	175
C.2	Algorithm for application of update rules	176

List of abbreviations

AI artificial intelligence

AUC area under the ROC (receiver operating characteristic) curve

BKOS framework for explanatory dialogue management; the name is an amalgamation of the word *because* and Ginzburg (2012)'s term KoS which relates to conversation-oriented semantics

BMI body mass index

CATE conditional average treatment effect

CBR case-based reasoning

CoT chain of thought

CXAI conversationally explainable AI

DAX deep argumentative explanations

DiCE diverse counterfactual explanations

ECG electrocardiography

GA global assessment of pain

GAM generalised additive model

GLM generalised linear model

ICM interactive communication management

IS information state

ISU information state update

LIME local interpretable model-agnostic explanations

LLM large language model

MAE mean absolute error

ML machine learning

NLG natural language generation

NLU natural language understanding

PDP partial dependence plot

QUD question under discussion

SCAM shape constrained additive model

SHAP Shapley additive explanations

XAI explainable artificial intelligence

Previously published material

Parts of this monograph are based on previously published material:

- Sections 3.4.2, 5.2, 5.3.1 and 5.4 are based on Alexander Berman, Eleni Gregoromichelaki and Catharina Parai (2025). ‘From Interpretability to Clinically Relevant Linguistic Explanations: The Case of Spinal Surgery Decision-Support’. In: *Proceedings of the 17th International Conference on Agents and Artificial Intelligence—Volume 1: IAI*. Presented in the Special Session on Interpretable Artificial Intelligence Through Glass-Box Models (IAI 2025) (<https://icaart.scitevents.org/IAI.aspx?y=2025>). INSTICC, Porto, Portugal: SciTePress, pp. 909–920. ISBN: 978-989-758-737-5. DOI: [10.5220/0013403800003890](https://doi.org/10.5220/0013403800003890).
- Chapter 4 is based on Alexander Berman and Christine Howes (2026). ‘Informing Design and Research Concerning Conversationally Explainable AI Systems by Collecting and Distilling Human Explanatory Dialogues’. In: *Information* 17.2. ISSN: 2078-2489. DOI: [10.3390/info17020123](https://doi.org/10.3390/info17020123). URL: <https://www.mdpi.com/2078-2489/17/2/123>.

CHAPTER 1

Introduction

In 2020, the Swedish public employment service introduced an artificial intelligence (AI) tool to assist decisions concerning access to employment support. After having been trained on historical data about previous jobseekers, a so called neural network predicts the probability that a newly registered jobseeker will find a job within 6 months, based on various information about the jobseeker. On the basis of this prediction, the tool recommends what kind of support the jobseeker should get (Berman, K. d. Fine Licht and Carlsson 2024; Berman 2024a). However, the caseworkers who were required to use the AI tool received very limited information about the basis of its recommendations. The tool merely lists 10 factors that influenced a specific assessment, e.g. the jobseeker's unemployment duration and city of residence, without explaining how those factors contributed to the AI's recommendation. In interviews, agency officials describe the system as a "black box" and that "you do not really know why it has come up with a certain decision". A caseworker recounts that one "cannot really figure out why one got that decision ... sometimes one cannot really understand what tilts the scale if you get a no for example".

This thesis is born out of a paradoxical situation. On the one hand, humans have created an immensely powerful technology that, in principle, can help us make better decisions faster, cheaper and with less effort. On the other hand, it is often difficult for us as humans to understand how this technology "thinks" and on what basis it makes its judgements and recommendations. Without such understanding, we cannot assess the reliability of its recommendations, be accountable for decisions that are made on the basis of its advise, or learn from the technology to advance our own knowledge.

To resolve this paradox, this thesis explores the possibility of combining the predictive and statistical power of AI with human communicative strategies. Long before the advent of AI, humans invented a tool for managing trust and accountability and for transferring knowledge. This tool offers a diverse set of tricks, norms and strategies for resolving ambiguities, coordinating social interaction and creating shared understanding. What if this tool, *dialogue*, could be incorporated into AI in ways that enable humans to more easily understand how AIs reason, and thereby to better reap the benefits of the technology?

1.1 Explainable AI in the era of large language models

The development of so called large language models (LLMs) and conversational AI assistants based on such models, including OpenAI’s ChatGPT, Google Gemini and Microsoft Copilot, have enabled new ways of using and interacting with AIs. If the Swedish employment service decided to replace their current tool, which uses traditional, “non-conversational” machine learning (ML), with an LLM, their tool would have been able to explain its decisions in natural language in the form of a dialogue with its human users. Instead of vaguely listing 10 factors, the LLM would, at least in principle, be able to answer any question that a caseworker or jobseeker might have regarding the basis on which its recommendation is made.

However, LLMs lack certain prerequisites for reliably explaining their own reasoning, as two U.S. lawyers experienced in a macabre episode in 2023. On a flight from El Salvador to New York, a passenger named Roberto was struck in the knee by a serving cart. Later, Roberto hired a law firm to file a lawsuit against the airline. The lawyers produced a legal brief to show precedent, listing six previous related cases where airlines had been sued for similar incidents (Weiser 2023). However, it later turned out that none of the cases compiled by the lawyers actually existed. In their research on the case, Roberto’s lawyers had used ChatGPT. When the lawyers asked whether a specific case that it referred to was real, it responded that upon “double-checking, I found that the case *Varghese v. China Southern Airlines ...* does indeed exist and can be found on legal research databases such as Westlaw and LexisNexis” (U.S. District Court for the Southern District of New York 2023). In some sense, the lawyers had done their job: rather than naively assuming that the information presented by the AI tool was valid, they critically interrogated ChatGPT about the evidence on which its outputs were based. They used the tricks and strategies of dialogue, and yet, the outcome of the interaction with the AI was a flagrant failure.

1.2 Adhering to human explanatory norms and strategies

This thesis studies how AI systems developed with the purpose of assisting high-stakes human decisions can become more explainable by adhering to norms and strategies of human explanatory dialogue. Conventional ML models generally cannot “explain themselves” in dialogue with humans. LLMs can, but not in ways that can be trusted. For these reasons, other approaches towards conversationally

explainable AI (CXAI) need to be sought. This thesis proposes and investigates a particular path towards this goal: systematically analysing human explanatory dialogues and emulating observed human strategies in conversational human–AI interfaces in ways that faithfully reflect the actual reasoning patterns of ML models.

1.3 Research questions

Through a combination of linguistic, cognitive, design-oriented, technical and philosophical perspectives, this thesis explores the prospect of conversationally explainable AI and addresses the following overarching research question: How can knowledge concerning the norms and strategies of human explanatory dialogue inform design and development of AI systems that enable users to interrogate and receive reliable explanations for predictions by AI models in the form of a dialogue?

The overarching research question is addressed through a progression of interconnected studies that target more specific sub-questions and where insights from studies feed into each other.

In order to identify research challenges associated with modelling human explanatory behaviours, chapters 3 and 4 empirically and exploratively investigate the communicative strategies that humans use when addressing the epistemic basis of their claims and judgements in dialogue in settings with and without AI assistance.

Based on the finding in chapters 3 and 4 that interlocutors often explain judgements argumentatively (as claims supported by premises) and enthymematically (by omitting one or more premises), chapter 5 formally analyses how arguments in the form of claims and evidence can be extracted from ML models and presented enthymematically in ways that invite accurate inferences concerning model’s reasoning patterns, resulting in the proposal of an argument extraction method for a particular class of ML models (generalised linear models). In order to study the effects of using the method in a real-world use case, chapter 5 also empirically assesses how professional users perceive, interpret and use explanations generated using the proposed method, as well as potential improvements that users would like to see.

Finally, chapter 6 builds on the argument extraction method proposed in chapter 5 to investigate how the explanatory dialogue strategies identified in chapters 3 and 4 can be captured in a formal model of dialogue management that theoretically accounts for the high-level cognitive mechanisms that enable an agent to engage in dialogues concerning epistemic bases of judgements and that

can also be used as a basis for implementing reliable CXAI systems.

1.4 Methodologies and theoretical frameworks

To address its research questions, the thesis builds on previous work concerning cooperative communication (Grice 1975), argumentation (Toulmin 2003), enthymemes (Breitholtz 2020) and information-state based dialogue management (Larsson 2002; Ginzburg 2012) and uses a combination of empirical, analytical and technical methodologies.

In chapters 3 and 4, human–human explanatory dialogues are analysed by rewriting them into analogous human–computer interactions following a dialogue distillation method (Jönsson and Dahlbäck 2000; Larsson, Santamarta and Jönsson 2000) which is performed iteratively and incrementally in tandem with dialogue modelling (Larman and Basili 2003).

The argument extraction method proposed in chapter 5 uses mathematical deduction to formally prove empirically and theoretically justified desiderata and empirically evaluates the proposed method using a participatory design approach (Sanders and Stappers 2008).

Chapter 6 formally models empirically observed dialogue strategies building on previous work on modelling dialogue management as information-state updates based on linear logic (Maraev, Bernardy and Ginzburg 2020).

1.5 Positioning the thesis in the research landscape

The work presented in this thesis relates to several different research themes and traditions:

- Research into **conversational AI**, **dialogue systems** and **chatbots** focus on design, implementation and evaluation of human–computer interfaces based on natural language dialogue (Chen et al. 2017; McTear 2020). Some work in this area has focused on **explanatory dialogue systems** (Winograd 1971; Shortliffe 1976; Moore and Paris 1993; Wahde and Virgolin 2023). This thesis focuses specifically on dialogue systems for explaining ML predictions (Slack et al. 2023; Mindlin, Beer et al. 2025; Berman and Larsson 2025), here termed *conversationally explainable AI* (CXAI).
- **Explainable AI** (XAI) focuses on technical approaches for making AI systems and the basis of their outputs more transparent and accessible for humans (Phillips et al. 2020; Guidotti et al. 2021; Gunning et al. 2021).

The work in this thesis relates primarily to **interpretable ML** which can be conceived as a sub-field of XAI that focuses on well-performing (accurate) ML models that make their predictions in ways that can easily be understood by humans (Rudin 2019; Rudin et al. 2022).

- **Human-centered AI** incorporates perspectives from cognitive psychology, design theory, and related disciplines to study how AI can be developed to address human needs and values, both generally (Shneiderman 2022; Capel and Brereton 2023) and in the specific context of explainability (D. Wang et al. 2019; Ehsan et al. 2021; Q. Vera Liao and Varshney 2022; Kim, Maathuis and Sent 2024). This thesis uses a human-centered approach by empirically grounding CXAI development in analyses of human language use, and by involving users via participatory design.
- **Trustworthy AI** focuses on compliance with laws and ethical norms such as transparency and fairness. The terms **responsible AI** (Dignum 2019; Agarwal and Mishra 2021; Mikalef et al. 2022) and **ethical AI** (Mittelstadt 2019; Siau and W. Wang 2020) are sometimes used to refer to similar considerations. Explainability is commonly conceived as a key aspect of trustworthiness (Floridi 2021; Kaur et al. 2022; Berman, K. d. Fine Licht and Carlsson 2024; K. d. Fine Licht and Folland 2025). The present thesis relates directly to trustworthy AI, primarily through its focus on explanatory reliability.

In relation to prior work concerning CXAI, the main novelty of the thesis lies in its combined focus on explanatory reliability (explanations provided for AI outputs should accurately reflect how the AI makes its predictions) and empirical grounding in human explanatory dialogues (AI systems should be able explain their outputs in similar ways as humans explain their judgements in dialogue).

1.6 Outline of thesis

Chapter 2 lays a philosophical and normative foundation for the subsequently presented studies by discussing the nature and function of explanations generally, why explanations can be crucial in the context of AI (both generally and in conversational form), and why knowledge concerning norms and strategies of human explanatory communication can be important for the design and development of conversationally explainable AI systems.

Chapters 3 and 4 collect and analyse human explanatory dialogues in two different kinds of contexts: expert explanations for medical judgements in clinical

settings without AI assistance and lay explanations in an experimental setting with AI assistance. The collected exchanges are analysed by rewriting them into the kind of interaction that would occur if the interlocutor acting as explainer was a computer instead of a human, and by identifying capabilities that a dialogue system needs to possess in order to adhere to the observed human communicative strategies.

Based on the insights from chapters 3 and 4, chapter 5 analyses how arguments in the form of claims and evidence can be extracted from a particular class of ML models, enabling an AI system to reliably represent the underpinnings of its own epistemic states and convey those underpinnings in ways that reflect the argumentative structure of observed human explanations. Chapter 6 then proposes how such representations can be utilised as resources in a formal model of dialogue management that accounts for the high-level cognitive mechanisms used by an agent engaging in dialogue concerning epistemic bases of complex judgements, regardless of whether the agent is a human or a computer.

Chapter 7 summarises the findings of the research and discusses limitations, future work and broader implications of the research.

CHAPTER 2

Philosophical and normative framework

This chapter lays a philosophical and normative foundation for the subsequent chapters by elaborating the notion of an explanation as such (section 2.1), the need for explanations in AI (section 2.2), normative aspects of explanatory communication (section 2.3) and the relation between explanations and arguments (section 2.4). Sections 2.5 and 2.6 then discuss how explanations and explanatory norms relate to different types of AI. Finally, section 2.7 concludes the chapter by outlining a research direction supported by the illuminated philosophical and normative considerations.

2.1 What is an explanation?

An explanation concerns the relation between an *explanandum* (pl. *explananda*), i.e. the event or circumstance to explain, and the *explanans* (pl. *explanantia*), i.e. the evidence or piece of knowledge that explains the explanandum (Hempel and Oppenheim 1948). According to the deductive-nomological model of explanation, which has been applied mainly in the natural sciences, an event or state can be explained with reference to specific circumstances (causes) and general laws of nature (Hempel and Oppenheim 1948). For example, if a toddler falls to the ground while trying to walk (explanandum), this can be explained with the fact that the toddler's center of mass moved outside the body's base of support (specific circumstance) and that any body that is not adequately supported and acted upon by gravity will accelerate towards the ground (general law).

While the natural sciences generally target *ontic* explananda, i.e. objective events or states in reality, the social sciences are typically concerned with other types of explananda such as agents' attitudes, beliefs and behaviours. This thesis focuses specifically on *epistemic* explananda such as beliefs, judgements and decisions. Epistemic states can potentially be explained in many different ways, depending on research tradition and choice of perspective. A cognitive psychologist may explain belief formation in terms of information processing (Kahneman 2011) or cognitive dissonance (Harmon-Jones 2019). For example, if Sam believes that smoking a cigarette after lunch is okay, the theory of cognitive dissonance might explain this belief as follows. Sam knows that smoking is

generally harmful, but smokes regularly. This causes a cognitive dissonance, which Sam resolves by focusing on the fact that her grandmother smoked all her life and lived to 90, or that smoking only once in a while is fine.

It is often conceived that explanations primarily serve the purpose of learning (Lombrozo 2006). Information concerning causes of events or reasons for actions help agents construct better models of the world that can be used for prediction and control (Heider 2013). Furthermore, explained decisions are more acceptable and increase trust in the decision-maker, compared to unexplained decisions (Lind, Kanfer and Earley 1990; Tyler 2006). In the context of public decision-making, explanations contribute to perceived legitimacy (K. d. Fine Licht and J. d. Fine Licht 2020) and are required by law in many countries in order to counter ambiguity concerning the true reasons behind a decision, to strengthen transparency and trust in the administration, and to facilitate review and legal contestation (Olsen, Slosser and Hildebrandt 2020). In collective decision-making processes, reason-giving can lead individuals to change opinion (Schneiderhan and Khan 2008) and is commonly conceived as fundamental in supporting legitimacy (Habermas 1984; Habermas 1990).

This thesis focuses on how agents communicatively explain the epistemic basis of their own judgements, and interrogate other agents' epistemic basis of their judgements. This perspective relates to the notion of folk psychology, which studies how people in everyday contexts try to understand and explain other individuals' behaviour (see e.g. Malle 2006), and follows in the tradition of studying explanations as communicative acts (Malle 2006, pp. 153–155). Scholars in this tradition have emphasised that explanations, like all contributions in dialogue, are constrained by general norms and principles of dialogue (Grice 1975; D. J. Hilton 1990) (see section 2.3), target questions raised by the explainee (Turnbull 1986), address the explainee's particular knowledge gap (Slugoski et al. 1993; Todorov, Lalljee and Hirst 2000; McClure and D. Hilton 1997) and allude to the explainee's background knowledge (Ducret 1988; Asher and Lascarides 2003; Breitholtz 2020). Furthermore, explanations are integrated with other social functions such as negotiation, argumentation, and impression management, thereby reflecting not only the explainer's causal view of the world (in the spirit of scientific explanations) but also social considerations of the explainer (Tedeschi and Reiss 1981; Harvey, Weber and Orbuch 1990; Charles Antaki and Leudar 1992; C. Antaki 1994). Hence, unlike the notion of explanation which is predominant in natural sciences, where an explanation reveals causes and regularities to explain events in the world, communicative studies of explanations also target social dimensions and functions of explanations in interpersonal relations.

2.2 Explanations in the context of AI

Many, if not all, of the functions of explanations mentioned above are potentially relevant in the context of AI and thereby serve as normative reasons for making AI explainable (T. Miller 2019; Mittelstadt, C. Russell and Wachter 2019; K. d. Fine Licht and J. d. Fine Licht 2020; Olsen, Slosser and Hildebrandt 2020; Vasconcelos et al. 2022; Grimmelikhuijsen 2023; Berman, K. d. Fine Licht and Carlsson 2024). For example, if AI is used in healthcare to predict individuals' risk of developing cancer, then the ability for medical professionals to learn from the AI can potentially benefit medical knowledge and thereby improve the quality of care. A doctor who is informed of the reasons for why an AI recommends a certain medical procedure for a particular patient can, at least in principle, more easily assess whether the AI's recommendation is reliable and potentially make a better decision by considering the evidence highlighted by the AI. A customer whose loan application is rejected on the basis of an AI's prediction is in a better position to contest the decision or obtain a positive decision in the future if the rejection is accompanied by a relevant explanation.

Similarly, emulating the interactive and conversational nature of human explanations can potentially be advantageous in the context of human–AI interaction (T. Miller 2019; Dazeley et al. 2021; Rohlfing et al. 2021; Booshehri, Buschmeier and Cimiano 2024). Perhaps most evidently, it is desirable that AI systems can answer the kind of specific explanation questions that users ask (Q Vera Liao, Gruen and S. Miller 2020) and support the kind of conversational strategies that humans use in explanatory dialogues (Mindlin, Beer et al. 2025; Berman and Larsson 2025). This need has been validated in interviews with decision-makers, where respondents indicate strong preferences for interactive explanations in the form of natural language dialogues over non-interactive explanations; specifically, respondents appreciate the prospect of a free-flowing dialogue and want to be able to formulate their own questions as well as get answers to follow-up questions (Lakkaraju, Slack et al. 2022).

2.3 Normative aspects of explanatory communication

Before discussing how knowledge about the conversational strategies used in human explanatory interactions can inform design of explainable AI systems (as will be elaborated in section 2.7), it is important to note that some social functions of explanations might not be desirable in the context of AI. For example, an AI system designed to be perceived as convincing may provide plausible explanations

that do not reflect the actual basis for its assessments. To illuminate this tension between different communicative ideals, we can turn to the cooperative principle laid out by Grice (1975). This principle states that explanations—like all kinds of contributions in conversation—should ideally follow certain communicative maxims or norms. According to the maxim of *relevance*, contributions should be relevant in their specific conversational context. The maxim of *quantity* dictates that one should say enough and not be overly verbose, while *quality* entails that one should say what one believes to be true and has evidence for. Finally, *manner* involves clarity of language, e.g. avoiding ambiguities and expressing oneself clearly.

While all of these maxims may seem self-evident from a normative perspective, adhering to them is not always easy in practice. Emma asks Noah why he thinks it will rain soon. Noah might think of several different kinds of responses that could potentially be considered relevant, e.g. referring to observable evidence (“dark clouds are approaching”), statistics (“it often rains this time of the year”) or an authoritative source (“my phone says so”). In terms of quantity, Noah can choose to mention many types of relevant information (evidence, statistics, source) and express himself at a high level of detail, or briefly mention only one of relevant pieces of information, depending on his view of Emma’s needs and interests. The maxims do not dictate exactly how Noah should behave, but serve as guiding principles.

Quality can be difficult to honour in situations where the explainer has limited insights concerning the basis of his own beliefs. Perhaps Noah has a strong feeling that rain is coming, without knowing exactly where this feeling comes from; his senses combine visual, auditory and barometric cues in ways that Noah is not fully aware of and cannot consciously introspect. The appropriate behavior in such a situation, considering the maxim of quality, is to abstain from answering the question due to lack of evidence (“I don’t know”) or possibly to mention the source of one’s belief in generic terms without making unsubstantiated detailed claims (“I just have a strong feeling”).

Sometimes, the norms of cooperative conversation are in tension with each other or clash with other social norms or conventions (Grice 1975, p. 47). A patient asks why the doctor recommends surgery; the doctor has made an intuitive judgement that surgery is the best course of action for this specific patient, but instead of referring to his intuition (“I just have a strong feeling”), which might seem inappropriate for a person in his social and professional role, the doctor explains his recommendation with reference to certified evidence (“studies show that surgery is most effective treatment for this disorder”). In some situations, epistemic transparency is less important than other communicative norms and functions, such as persuasion or coordination of trust (Mercier and Sperber 2017).

Which norms should AI systems prioritise? The answer to this question depends on the intended purpose and function of its application. In low-stakes scenarios such as tutoring or movie recommendations, it might be appropriate for AIs to justify their predictions and recommendations as convincingly as possible, rather than transparently explain their reasoning process. For example, if a language learner wants to know why an AI states that a particular idiom is preferred over another one, a convincing argument might be more useful than a detailed explanation of how the AI reasoned. However, in the kinds of scenarios targeted in this thesis, where AI systems are intended to assist high-stakes decisions, epistemic transparency can be assumed to be more important than argumentative vigilance.

2.4 Arguments and other related concepts

The distinction between explanations and related concepts such as arguments, reasons, justifications and rationales is not clear-cut. As already discussed, in social life, explanations often serve multiple purposes. Similarly, in the context of the research presented in this thesis, these concepts largely overlap and cannot easily be kept apart. Nevertheless, a distinction will here be upheld between explanations and arguments. Following Bex and Walton (2016, p. 62), we can say that an explanation serves to facilitate the addressee's understanding (in this case concerning the speaker's epistemic basis of a judgement), while an argument primarily serves to convince the addressee of some unsettled claim.

Communicatively, the close relationship between explanations and related concepts constitutes a source of potential ambiguity. If a doctor recommends that the patient should undergo surgery "since studies show that surgery is the most effective treatment for this disorder", this can be understood either as an explanation concerning the epistemic basis of the doctor's judgement, or as an argument/reason/justification for his claim—or both. In many situations, potential differences between these interpretations may be negligible, especially when the receiver perceives the sender as trustworthy and finds the claim and explanation/argument reasonable. However, in other situations, ambiguities can be more problematic. As an illustration, we can conceive of a scenario where the social insurance agency rejects Oliver's application for sickness benefit on the stated basis that Oliver, although physically hindered to perform his present work, can seek other jobs for which his medical condition does not constitute an obstacle. We also imagine that in reality, the caseworker processing the application has another basis for rejecting the application: since Oliver lives in a neighborhood where welfare fraud has repeatedly been reported, the caseworker suspects that

Oliver is simulating his symptoms. This is the actual reason behind the decision. In other words, the stated reason—which can be interpreted either as an explanation, argument or both—is in fact only an argument. The hypothetical scenario illustrates that potential ambiguities concerning whether a communicative act is an explanation or argument/justification can violate the maxim of manner and that such violations are particularly impactful in situations where explanations are intended to enable transparent decision-making, such as in public administration (K. d. Fine Licht and J. d. Fine Licht 2020; Olsen, Slosser and Hildebrandt 2020). In an AI context, potential negative consequences of such ambiguity problems can be avoided by ensuring that provided explanations are *faithful* with respect to how the AI actually reasons, as will be elaborated below.

2.5 Different types of AI explanations

In order to understand how notions like explanation and epistemic transparency play out in the context of AI, it is crucial to distinguish between different types of AI. In broad terms, one can distinguish between AI systems that are *explained by a human–AI interface* and *self-explaining AIs*. Historically, the former category has been most common, while the latter category has emerged only in recent years.

Early AI incarnations, often referred to as expert systems or symbolic AI, were primarily based on hand-crafted rules which enabled them to draw conclusions from specific pieces of information (Gunning et al. 2021). For example, the expert system MYCIN used rules to identify bacteria causing infections and to recommend antibiotics (Shortliffe 1976; Buchanan and Shortliffe 1984). The system traced its own reasoning and could explain a specific assessment by referring to the rules that it had used to reach its conclusion. Conceptually, this behaviour can be characterised as a mechanistic explanation, since it conveys the process behind the explanandum (Lombrozo and Gwynne 2014).

The early generation of AI systems were generally ineffective, since the relatively crude rules based on which their inferences were drawn failed to capture the complexities of the real world. In contrast, more recent AI systems are typically based on ML, which has enabled more accurate predictions (LeCun, Bengio and Hinton 2015; Lappin 2021; Gunning et al. 2021). Within this paradigm, a model acquires certain aspects of its reasoning by being trained on examples (supervised learning) or with feedback (reinforcement learning) (S. Russell and Norvig 2022). For example, a bank might train a model to predict whether a new loan applicant will be able to pay interest and mortgage, based on historical data concerning previous clients. Many ML models, such as neural networks and random forests, combine different kinds of information in intricate ways, making the basis for

specific predictions difficult to identify and communicate. Notably, mechanistic explanations for ML predictions are not very helpful, since the process leading up to a specific prediction might be difficult to understand for humans, even for the individuals that developed and trained the AI. Consequently, such ML models are often described as “opaque” or “black boxes” (Burrell 2016; Rudin 2019).

When the term explainable artificial intelligence (XAI) was popularised in 2017 through the launch of a DARPA project, the focus was explicitly on ML explanations (Gunning et al. 2021). Since then, various methods and approaches have been developed and proposed for making ML models more explainable. These approaches roughly fall into two separate research directions: development of methods for extracting explanatory information from opaque models, and development of ML models that are inherently interpretable by design (Rudin 2019; Gunning et al. 2021).

Regardless of whether an AI is based on hand-crafted rules or ML, claims and corresponding explanations are communicated to users through an interface, e.g. a dialogue system or an automatically generated textual explanation. The development of human–AI interfaces involves design choices that substantially shape how claims and explanations are framed when communicated to users. For example, the Swedish public employment service uses an ML model to estimate jobseekers’ need of support, and explains decisions concerning employment support with written statements such as: “In your case, it was primarily the following factors that contributed to the assessment: your unemployment duration, your unemployment history, your city of residence, and working time.” (Berman 2024a) Notably, such statements are generated (automatically) using a template developed by professionals at the employment service, rather than by the AI. In this sense, neither rule- or ML-based AIs “explain themselves”; instead, their outputs are interpreted and explained by humans, e.g. on the basis of knowledge about the AIs mode of reasoning or other explanatory information extracted from the AI.

With the more recent advent of LLMs and so called generative AI, the notion of explainability takes on a different form. LLMs are trained on massive amounts of content from books, web pages and other sources, primarily on the task of predicting the next token¹ in a piece of text (Vaswani et al. 2017; Brown et al. 2020). Through such training, LLMs acquire a broad range of linguistic and non-linguistic capabilities that can be used for a wide variety of tasks, including for decision-support. Interestingly, one of the capabilities that emerges from the conceptually simple task of predicting text is to conversationally explain the AI’s own claims and judgements. Conceptually, this radically distinguishes LLMs from

¹Tokens roughly correspond to words.

rule-based and conventional ML in that LLMs, in some sense, explain themselves.

As will be discussed below, the extent to which AI systems can adhere to normative principles of explanatory communication hinge on the differences between types of AI outlined above.

2.6 Prerequisites for epistemically transparent AI

To be able explain the epistemic basis of one’s beliefs in ways that honour the maxim of quality, an agent requires knowledge concerning the epistemic underpinnings on which its beliefs are formed. This requirement is not always fulfilled for humans (Polanyi 1966; Wason and Evans 1974; Nisbett and Wilson 1977; Khaneman 2003; Messner 2009; Mercier and Sperber 2017) and can also be difficult to satisfy for AI systems, depending on the type of technology.

For AI systems that are explained with an interface developed by humans, epistemic transparency hinges on two capacities: (1) obtaining faithful representations of the underpinnings of AI outputs and (2) communicating the content of such representations adequately. The technical notion of explanatory *faithfulness* concerns the extent to which explanatory information concerning an AI’s output reflects the AI’s actual reasoning.² Faithfulness differs substantially between different types of AI. For rule-based AI systems and inherently interpretable ML models, faithful representations can be obtained by automatically tracing the system’s inferences (Shortliffe 1976; Buchanan and Shortliffe 1984) or leveraging knowledge about the AI’s actual reasoning steps (Rudin 2019; Molnar 2025). For example, a sparse decision tree predicts outcomes by following a sequence of branching choices that can be communicated to users, while predictions by linear models can be accurately explained in ways that resemble how humans frequently express arguments (as will be shown in chapter 5).

Faithful representations can be obtained for opaque ML models as well, by communicating the entire sequence of computational operations between input and output. For example, if a bank assesses creditworthiness using a random forest classifier (which consists of a large number of decision trees), a faithful explanation might be expressed as follows:³

The application is rejected since most (88 out of 100) of the model’s decision trees predict **no**, i.e. that the applicant will not be able to pay interest rates and mortgages in the future. Specifically, **tree 1** predicts

²See e.g. Rudin (2019), Jacovi and Goldberg (2020), Lanham et al. (2023) and Turpin et al. (2023). Phillips et al. (2020) use the term *explanation accuracy* while Guidotti et al. (2021) use the term *fidelity*.

³The example is a hypothetical explanation for an actual classification by a random forest classifier trained on the German credit dataset (Hofmann 1994).

no because the applicant has no checking account, a credit amount of 8,724 or less, a personal status of female (divorced, dependent, or married), a credit history where previous loans were not all fully paid, employment between 4 and 7 years, a loan duration greater than 16 months, no known savings, no other payment plans at stores, a non-business loan purpose, savings between 500 and 1,000 Deutsche Marks, owns their housing, no checking balance of 200 marks or more, and is 32 years old or younger. **Tree 2** predicts **no** because ... (etc.)

As indicated by the hypothetical explanation above, faithful (mechanistic) explanations for opaque models tend to be overly detailed and verbose (hence violating the maxims of quantity/manner), without exerting much explanatory force (hence failing with respect to relevance). In practice, when opaque ML models are used in real-world applications, they therefore tend to be explained in more approximate terms and in ways that sacrifice faithfulness for (some amount of) intelligibility. For example, the relative importance of variables (features) can be estimated and presented to users (Ribeiro, Singh and Guestrin 2016; Lundberg and Lee 2017); this is how the Swedish public employment service explains decisions concerning allocation of employment support (Berman 2024a). However, since estimates of feature importance are approximate, they can be inaccurate (Amparore, Perotti and Bajardi 2021; Krishna et al. 2025). Furthermore, approximate explanations of these kinds are not necessarily perceived as intelligible by users (Berman, K. d. Fine Licht and Carlsson 2024), for reasons which will be discussed in chapter 5.

As for self-explaining LLMs, the distinction between explanatory representations and communication with user cannot easily be upheld; in fact, the opacity of such models makes it difficult to study whether LLMs even have stable internal representations.⁴ Instead, when epistemic transparency is assessed, it is typically done experimentally, similar to psychological studies of humans (cf. Nisbett and Wilson 1977). In this regard, one can distinguish between *retrospective* explanations and so called *chain of thought (CoT) reasoning* (Wei et al. 2022). The anecdote in section 1.1 exemplifies the former kind: First, the LLM makes a claim (produces a list of six legal cases related to the case researched by the user). Then, when the user asks for evidence, the LLM produces a retrospective explanation (that it found the cases in legal databases). As LLMs generate their outputs by producing one token at a time, without maintaining an internal state (Vaswani et al. 2017; Brown et al. 2020), it is easy to understand why retrospective LLM explanations can be unreliable. Without keeping track of the inferences upon which a claim is initially made, an LLM cannot reliably explain those inferences in retrospect.

⁴See e.g. Uceda-Sosa et al. (2024) and Rai et al. (2025).

Instead of revealing the actual basis on which its claim was made, the LLM reconstructs (or confabulates/hallucinates) a reason that seems plausible and that potentially serve as a valid argument for the claim, but not necessarily a reliable explanation for it.

In principle, CoT reasoning can yield more reliable explanations, since the LLM generates its claim at the end of a reasoning chain that can be presented to the user as an explanation for the claim. For example, in its “thinking” mode, ChatGPT 5 can reveal a trace of its reasoning, including information retrieval from external sources (OpenAI 2025). Nevertheless, empirical studies demonstrate that CoT explanations are not always faithful. Turpin et al. (2023) found that LLM assessments are frequently associated with stereotypes related to categories such as race/ethnicity and gender identity, but that the influence of such stereotypes on model assessments was not mentioned in their CoT. Lanham et al. (2023) experimentally manipulated chains-of-thought and observed that LLMs’ conclusions do not always depend on the CoT. Together, these studies indicate that neither retrospective nor CoT explanations LLM are generally reliable.

2.7 Choice of research direction

As this chapter has aimed to elucidate, research into CXAI systems can target different intended purposes of AI use and be guided by distinct normative ideals. The path taken in this thesis is to explore how the predictive power of ML, which potentially enables better decision-making, can be leveraged while adhering to human conversational norms and strategies, particularly concerning epistemic transparency. The specific focus on epistemic transparency is intended to be aligned with principles of trustworthy AI usage (Floridi 2021; Kaur et al. 2022; Berman, K. d. Fine Licht and Carlsson 2024; K. d. Fine Licht and Folland 2025).

From a technical perspective, the emphasis on accurate decision-making favours focusing on ML models over hand-crafted rule-based AI for predictions. The emphasis on epistemic transparency disfavours self-explaining LLMs and favours development of dialogue systems serving as interfaces to inherently interpretable ML models, since such systems can, at least in principle, be designed to explain predictions faithfully. It is within this conceptual and technical framework that the subsequent chapters empirically study conversational strategies in human explanatory dialogues as a basis for informing design and implementation of explanatory dialogue systems acting as interfaces between human users and ML models.

In order to empirically ground CXAI systems in human conversational behaviours, chapters 3 and 4 will collect and analyse human explanatory dialogues

with the purpose of identifying capabilities that such systems need to possess in order to emulate observed human communicative strategies. The empirical grounding in human language use is motivated by a human-centered approach to AI development which focuses on actual needs, demands and behaviours of human users (Shneiderman 2022; Capel and Brereton 2023). While previous work has empirically investigated users' needs for conversational AI explanations (Lakkaraju, Slack et al. 2022), the extent to which different kinds of dialogue moves contribute to explanatory success (Booshehri, Buschmeier and Cimiano 2024), the kind of questions that human users might want to ask to an XAI system (Q Vera Liao, Gruen and S. Miller 2020), as well as users' interactions with existing CXAI systems (Sokol and Flach 2020; Slack et al. 2023; Wijekoon et al. 2024; Mindlin, Robrecht et al. 2024; Mindlin, Booshehri and Cimiano 2026), in terms of concrete dialogue strategies, existing CXAI approaches tend to be based on the researcher's (sometimes implicit) assumptions concerning the kinds of behaviours and capabilities that it would be useful for such systems to exhibit and possess. An evident drawback of such an approach is that the researcher's assumptions may not properly reflect actual communicative behaviours of human explainers or explainees, which might limit the potential usefulness and value of systems designed on the basis of such assumptions. It is for these reasons that the proposed approach collects and exploratively analyses human explanatory dialogues with minimal prior assumptions, as will be further elaborated in the subsequent chapter.

CHAPTER 3

Explanations in non-AI-assisted medical dialogues

In order to identify the kind of dialogue strategies that a CXAI system might need to possess, as well as research challenges associated with modelling such strategies, this chapter collects and analyses epistemic explanations in medical dialogues in clinical settings. The medical domain is chosen since explanations can be assumed to have a central role in contexts involving high-stakes decisions where it is important to uphold trust (Lind, Kanfer and Earley 1990; Tyler 2006; Van Eemeren, Garssen and Labrie 2021).

Three empirical sources of clinical spoken dialogues were chosen: the Norwegian corpus of doctor–patient consultations from Ahus (Gulbrandsen, Finset and Jensen 2013, henceforth abbreviated Ahus), a Swedish textbook in medicine focusing on the encounter between patient and doctor (Lindgren and Aspegren 2004, henceforth abbreviated L&A), and American anesthesiologists’ interactions during neurosurgical operations (Johannesen, R. I. Cook and Woods 1994, henceforth abbreviated Johannesen). The choice of empirical material is primarily motivated by the topics and types of situations that it encompasses. Furthermore, while two of the corpora (Ahus and Johannesen) are descriptive and contain transcripts of actual consultations and expert interactions, the other (L&A) is prescriptive and conveys communicative norms. As manifestations of human explanatory strategies in high-stakes settings, both types of linguistic data were deemed relevant for the purposes of the research.

Collected dialogue excerpts are analysed by rewriting them to fully implementable analogous human–computer interactions, using a method for dialogue distillation described in section 3.3. The analysis is explorative in that it aims to identify research challenges associated with modelling observed human explanatory strategies, with a minimal set of a priori assumptions concerning the nature of the collected data (e.g. types of illocutions, questions, etc.), thereby enabling an open-ended investigation.

3.1 Related work

Previous work has collected and/or analysed explanatory dialogues both in health-care settings (Peräkylä 1997; Heritage and Maynard 2006; den Boeft et al. 2017;

Josephine B. Fisher et al. 2022) and other domains (Moore and Paris 1993; Madumal et al. 2019; Attari, Heckmann and Schlangen 2019; Booshehri, Buschmeier and Cimiano 2024; Josephine B Fisher et al. 2023; Götze and Schlangen 2023; Buhl, Josephine B. Fisher and Rohlfing 2024). While these prior works have yielded insights that could potentially inform design and implementation of CXAI systems in some regards, no previous work has specifically analysed human explanatory dialogues in relation to ML predictions or statistical inferences with the purpose of informing design and development of CXAI systems. For example, the explanation dialogue model derived by Madumal et al. (2019) from six different types of human explanation dialogues offers an abstract formalisation of how the state of an explanation dialogue can be updated via different types of locutions (explain, affirm etc.), but does not account for dialogue phenomena specific to ML predictions or statistical inferences.

To address this gap in previous research, the present chapter uses methodological insights from empirically informed dialogue system development and applies them to XAI. One established method for empirically grounding dialogue system development in analogous human interactions is to collect relevant empirical data by letting participants interact with a system operated by a so-called Wizard of Oz, i.e. a human operator controlling the system behind the scenes (Dahlbäck, Jönsson and Ahrenberg 1993). Typically, the wizard chooses system responses from pre-defined canned texts and templates, thereby enabling swift and machine-like responses. One disadvantage of Wizard-of-Oz approaches is that they might not realistically reflect how humans naturally interact with each other (Jönsson and Dahlbäck 2000). This has motivated the development of an alternative and complementary approach called dialogue distillation (Jönsson and Dahlbäck 2000; Larsson, Santamarta and Jönsson 2000), where human-human dialogues are rewritten into the kind of dialogue that would occur if one of the interlocutors was a computer. The present chapter uses (and extends) this approach to analyse explanatory strategies in medical dialogues. While Diana C Hernandez-Bocanegra and Ziegler (2021) used a Wizard-of-Oz approach to study the kinds of explanatory questions that users might ask an explainable recommender system, as far as the author is aware, no previous work has applied empirical dialogue system development methods to inform design and development of CXAI systems.

3.2 Data

Occurrences of explanations (or related phenomena such as arguments or justifications) pertaining to medical judgements were identified using a search procedure. In the case of Ahus, which contains transcriptions of 220 consultations, this was

done by searching for the word “why” (“hvorfor” in Norwegian) to identify dialogues of potential interest and then manually searching for explanations within the dialogues.¹ It should be noted that the search procedure does not result in a random sample of the corpus, and fails to discover explanatory exchanges that do not contain the word “why”. For the other empirical sources, the material was small enough to permit a manual search of the entire data.

The topic of interest (medical judgements) primarily encompasses diagnosis (judging that a patient has a certain condition) and recommendations (judging that a particular action or intervention is adequate). The selection procedure resulted in 12 dialogues (Ahus: 4, L&A: 2, Johannesen: 6), spanning a total of 130 utterances.

3.3 Method

Building on a method for dialogue distillation proposed by Jönsson and Dahlbäck (2000) and Larsson, Santamarta and Jönsson (2000), collected excerpts of human–human dialogues are manually rewritten into analogous human–computer interactions (as will be exemplified in section 3.4). The purpose of dialogue distillation is to inform dialogue system development by rewriting human–human dialogues into the kind of interactions that would occur if one of the interlocutors, in this case the explainer, was a dialogue system. In addition to a corpus of rewritten human–computer dialogues which can be used to inform design and modelling of dialogue systems, dialogue distillation also yields insights concerning the kind of dialogue capabilities that a dialogue system would need to have in order to emulate the role of the “replaced” human.

The distillation process generally follows a protocol (or set of guidelines) and builds on certain assumptions that may differ between domains and applications as well as the specific aims of the analysis. Previous work on dialogue distillation (Jönsson and Dahlbäck 2000; Larsson, Santamarta and Jönsson 2000) assumes a waterfall approach where the protocol is developed before rewriting dialogues, and where the dialogue system is implemented in a later stage of development to support the behaviours manifested in the rewritten dialogues. One limitation of this approach is that it can be difficult to anticipate a priori the kind of methodological choices that the protocol should inform, without first attempting to rewrite (at least some of the) dialogues. Furthermore, if dialogues are rewritten before system implementation, this can potentially result in a corpus with formally inconsistent

¹The Ahus corpus contains multilogues, involving not only doctors and patients, but also other healthcare staff and relatives. The present analysis focuses only on contributions by doctors and patients.

or otherwise unimplementable dialogues. To mitigate these drawbacks, the present work adopts an iterative and incremental approach (Larman and Basili 2003) where distillation is performed in parallel with modelling of dialogue management so that, at any given iteration, the system capabilities manifested by the rewritten dialogues are supported by the dialogue model. Specifically, the iterative process begins with minimal initial assumptions (see below) that are refined across iterations, similar to the use of open coding in grounded theory (Strauss and Corbin 1990). This process is steered towards gradually increasing the coverage of the dialogue management model and decreasing the differences between the rewritten and original dialogues. In the present chapter, this process continues until a substantial amount of the observed phenomena are deemed to be adequately modeled; at this stage, any remaining salient differences between original and rewritten dialogues can be considered challenges for future work (as will be elaborated in section 6.5).

The dialogue distillation was performed by the author, partly on the basis of feedback from the author’s main supervisor who reviewed in-progress distillation results.

3.3.1 Technical assumptions

The dialogue system acting as substitute for the human explainer is assumed to have general (domain- and language-independent) dialogue management capabilities as well as domain-specific resources for natural language understanding and generation (NLU, NLG) and for making inferences, e.g. using predictive models. Human interlocutors’ judgements are assumed to potentially correspond to model predictions. For example, if a doctor, based on available evidence and circumstances, expresses that the patient probably has diabetes, then this judgement is assumed analogous with the prediction of a model trained to classify whether a patient has diabetes.

The dialogue system is assumed to use some kind of information-state based dialogue management (Larsson 2002; Ginzburg 2012; Maraev, Bernardy and Ginzburg 2020), where the system keeps track of the state of the dialogue and selects which dialogue moves (speech acts) to perform by iteratively applying update rules. Importantly, however, no initial assumptions are made concerning the range of potential moves handled by the system, the structure of the system’s information state, or the system’s update rules. Instead, the elements of the system are incrementally developed during the distillation process. The final resulting dialogue model, based on the distillation in the present chapter and chapter 4, will be presented in chapter 6.

3.3.2 Normative assumptions

In addition to the theoretical goal of modelling the human–human dialogues as faithfully as possible, it is also assumed during distillation that the dialogue system should facilitate for users in forming *accurate mental models* of the basis on which the system’s predictions are made. This assumption reflects normative principles relating to AI transparency and trustworthiness (Berman, K. d. Fine Licht and Carlsson 2024) and concretely implies that explanations delivered by the dialogue system for its predictions should reflect the actual inner logic of the predictive model at hand. While this requirement may seem normatively evident, its implications for design and modelling of explanatory dialogue systems are not trivial, at least in cases where human interlocutors explain judgements using types of information that cannot straightforwardly be derived from a predictive model (as will be discussed in section 3.4.9).

3.3.3 Iterative and incremental process

To illustrate the iterative and incremental nature of the distillation process, we can consider the excerpt below, where the doctor (D) unpromptedly provides three reasons for switching medication:

- (3.1) D: and then I would like uh to change , since you have had over two months and have increased the dosage and not had any effect , so I want to suggest another uh antidepressant medication called Effexor (Ahus)

In early iterations, the dialogue management model only supported providing one piece of evidence or explanation at a time, and only when requested by the user. Consequently, excerpt 3.1 above was rewritten as below (s=system, u=user):

- s: I suggest switching medication.
u: Why?
s: You have had your current medication for a relatively long time.

In later iterations, the dialogue model was extended to support multiple pieces of evidence and unprompted explanations, causing the distilled excerpt to be refined as follows, achieving a higher degree of coverage:

- s: I suggest switching medication since you have had your current medication for a relatively long time and have increased dosage and not had any effect.

The results presented in the next section pertain to the final stage of the distillation process.

3.4 Results

The distillation process revealed a broad range of dialogue phenomena and strategies which might be relevant in the pursuit of human-aligned conversationally explainable AI systems, as will be presented in the subsequent sections. (Issues concerning how to formally model these behaviours are discussed in chapter 6.)

Cited excerpts from Ahus and L&A have been translated to English by the author; for original corpus material, see appendix A. Automated tests demonstrating that the rewritten dialogues are implementable using the dialogue model outlined in chapter 6 are available in the repository for the dialogue manager (Berman 2025a).

3.4.1 Explanation triggers and query types

Explicit requests for epistemic explanations are rare in the empirical material. One exception is a patient’s why-question in response to the doctor’s recommended action (D=doctor, P=patient):²

- (3.2) D: I just feel like taking an X-ray of your lungs as well, on the way up,
P: *why is that?*
D: no, just, because I looked at the oxygen percentage in your blood, which is a bit lower than one would expect,
P: *oh? how high is it?*
D: uh it’s at ninety-three percent, and one would expect between ninety-five yes, I mean it’s a bit unspecific,
...
P: *what does it mean to have too low oxygen uptake then? what could it be a sign of? that lung function is poor?*
D: uh in itself, it’s terribly unspecific, I just want to see that there isn’t anything obvious in the lungs (...) (Ahus)

Excerpt 3.2 also illustrates how explanation requests (second and third emphases) can occur in response to a previously delivered explanation. Specifically,

²Self-repairs and non-lexical symbols (e.g., pause markers and non-verbal actions) have been removed, unless if deemed important for the analysis.

D's explanation that P's oxygen percentage is low leads P to interrogate its specific level, and what the low oxygen saturation "mean[s]" and is a "sign of".

The excerpt below contains another example of question-begging explanations:³

- (3.3) P: [do you think] I can have these lungs replaced?
...
D: no, I suppose that, unfortunately, it isn't (feasible)
D: it rarely is , once in a while, it can be done, and it's extremely rare
...
P: *yes, why do you say that once in a while, it can be done, but not*
D: no, so uh in special types patients uh it is usually completely different kinds of diseases that we are talking about then (...)
(Ahus)

In excerpt 3.3, P's emphasised question can be understood as a request for an elaboration of D's explanation that lung transplants can be done only once in a while.

Occasionally, when a doctor provides evidence without making any claim, the patient asks a question that leads the doctor to express a claim supported by the evidence. Such questions can be understood as requests for a forward expansion (Xydis et al. 2020). In the excerpt below, P proposes a forward expansion and asks D to confirm it:⁴

- (3.4) D: Also, I want to inform you that your blood tests show your blood sugar is perfectly normal, and your cholesterol levels are good too.
P: *So I'm fairly healthy for my age?*
D: Yes, I'd say you are. Except for your high blood pressure, I mean your blood pressure condition, you have that (...)
(L&A, p. 17)

A similar pattern can be observed in the excerpt below:

³Minor contributions by two relatives of the patient have been excluded.

⁴Excerpt 3.4 occurs in a context where D has previously proclaimed that P's coughing is a side effect of medication against high blood pressure and hence not dangerous. P has not explicitly asked about the results of the blood tests or his/her overall health.

- (3.5) D: The sample taken during the bronchoscopy showed cellular changes.
 P: *What does the doctor mean by cellular changes?*
 D: It's a tumor.
 P: *Does the doctor mean that I have cancer?*
 D: Yes, you have cancer. (L&A, p. 251)

In excerpt 3.5, D's first utterance can be understood as a datum supporting a not-yet-expressed claim, which is delivered in response to P's first question. P's second question proposes an additional forward expansion (from the claim recently delivered by D), which is then confirmed by D. Argumentatively, D's reasoning can be understood as a chain of inferences (see section 3.4.5), where P drives the conversation "forward" from evidence to conclusions by asking questions to D.

In contrast to the examples above, in most cases, explanations are delivered unpromptedly, i.e., without evident conversational triggers, and in direct association with the explanandum, as exemplified by excerpt 3.1 (repeated below for convenience):

- D: and then I would like uh to change , since you have had over two months and have increased the dosage and not had any effect , so I want to suggest another uh antidepressant medication called Effexor

In excerpt 3.1, D provides an explanans (having had a medication for over two months and increased the dosage and not had any effect) in the same turn as the corresponding explanandum (the suggestion to switch antidepressant medication). In this case, the relation between the explanans and explanandum is linguistically marked with the discourse particle "since". In other cases, such relations are tacit, as in the excerpt below:

- (3.6) R: Have a feeling it's [hematocrit measurement] still gonna be low, he's just oozing [bleeding] all over the place
 (Johannesen, p. 36)

In excerpt 3.6, the rhetorical relation between the resident⁵ (R) anesthesiologist's prediction (that the hematocrit measurement will be low) and corresponding explanans (the patient's severe bleeding) is not marked linguistically, and needs

⁵A resident anesthesiologist is gaining practical experience for four years after medical school. An attending anesthesiologist is a senior member of the anesthesiology staff, holding a faculty position (Johannesen, R. I. Cook and Woods 1994).

to be identified pragmatically. A similar strategy is observed in the excerpt below, where the explanandum (an anesthesiologist's (A) assessment that the patient needs fluid) is tacitly linked to the explanans (that the patient's urine is dark):

- (3.7) A: ... He probably needs some fluid I would think, his urine
looks pretty dark (Johannesen, p. 39)

Due to the lack of evident conversational triggers, it is assumed during distillation that unprompted explanations of the kinds exemplified above are caused by extra-linguistic factors. For example, during surgery, a critical event such as an abnormal sensor reading might lead an anesthesiologist to recommend a particular course of action, and to justify that action. However, the role of such extra-linguistic factors is not modeled as such.

3.4.2 Types of explanantia

Interlocutors frequently explain beliefs and recommendations with reference to up to three specific circumstances. Following Toulmin (2003), concrete pieces of evidence in support of claims are here referred to as *data*. In doctor–doctor interactions (Johannesen), data are frequently conveyed with high numerical precision and in ways that assumes that the hearer understands how the information should be interpreted and used in the context, as in the excerpt below:

- (3.8) R2: *his pressure's now reading 177*
...
R: cuff can't read his pressure well. Give him another 50
mics of fentanyl. *I think it's a true pressure*
(Johannesen, p. 32)

In excerpt 3.8, R motivates the administration of fentanyl with reference to the fact that the patient's blood pressure is 177. Since R2 is a peer expert, R does not need to explicitly say that a blood pressure of 177 is abnormally high. In contrast, in doctor–patient consultations (Ahus and L&A), data are frequently conveyed in numerically vague but pragmatically informative ways. This strategy directs the patient's attention to the doctor's medical interpretation of the objective facts at hand. Presumably, such a strategy helps hearers identify what kind of warrant (general rule or principle; see Toulmin 2003) the speaker might have in mind. For example, when the doctor in excerpt 3.2 describes the patient's oxygen levels in the blood as “a bit lower than one would expect”, the words “lower” and “expect” allude at a warrant such as “unexpectedly low levels of oxygen in the blood can indicate a lung abnormality”. Similarly, when D in excerpt 3.1 explains a recommended change of medication with reference to having had the

medication for “over two months and ... increased the dosage and not had any effect”, the lexical choices “over”, “increased”, and “not ... any” trigger a warrant such as “having used a medication for a long time without any effects motivates trying another medication”.⁶

Generally, two types of warrant triggers can be observed: scalar/gradable and norm/expectation related. Examples of scalar triggers include *lower* (levels of oxygen in the blood than one would expect), *high* (blood pressure), *over* (two months of medication use), *increased* (dosage), *no* (effect of medication), *big* (airway), *pretty* (patient’s urine looks pretty dark), and *small* (nares). Examples of norm-related warrant triggers include *expect* (regarding oxygen saturation), *normal* (lung X-ray), *abnormal* (nothing abnormal in patient’s lungs), *should* (nothing observed that shouldn’t be there), *just fine* (heartbeat, ECG), and *good* (cholesterol levels, “other parameters”).

It is worth noting that in cases when a claim is supported by a relative (numerically vague) datum, it might later be elaborated quantitatively, as in excerpt 3.2 (repeated in part for convenience below):

- D: I just feel like taking an X-ray of your lungs as well, on the way up,
P: why is that?
D: no, just, because I looked at the oxygen percentage in your blood,
which is a bit lower than one would expect,
P: *oh? how high is it?*
D: uh *it’s at ninety-three percent*, and one would expect between ninety-five
yes, I mean it’s a bit unspecific,
...

Warrants are not only implicated with linguistic triggers, but also conveyed more explicitly. In the excerpt below, the attending (A) and resident (R) anesthesiologist discuss whether the patient should be given fiberoptic intubation orally or nasally:

- (3.9) R: Let’s just try an oral for once, see if we can do it
...
A: When he’s asleep I’ll tell you why I didn’t do an oral
...
A: *I just think with orals you really have to sedate him a lot more...*
R: See I don’t. *I’ve done a lot of them, you don’t have to sedate them at all* (Johannesen, p. 41)

⁶These examples also illustrate that warrants can be triggered by combining relative (e.g. “over”) and numerically more precise (“two months”) information (see section 6.3.7).

In excerpt 3.9, A's emphasised explanation can be interpreted as partly conveying a warrant, namely that oral intubation typically requires more sedation. Presumably, the plural form "orals" and the generalising pronoun "you" indicate a general scope (warrant), while the singular pronoun "him" links the warrant to the claim at hand. R's response constructs a counter-warrant (oral intubation frequently requires no sedation), which is backed by personal experience ("I've done a lot of them"). (In both cases, there is also an implicit inference step according to which methods that require less sedation are considered more favorable; see section 3.4.5.)

A similar warrant-based explanation strategy can be observed in the excerpt below, where D consistently explains his assessment (that a lung transplant is infeasible) with reference to general knowledge (about lung transplants):

- (3.10) P: [do you think] I can have these lungs replaced?
 D: [yes, that would have been] really nice, that
 D: hm ?
 P: it's not feasible
 D: no, I suppose that, unfortunately, it isn't
 D: *it rarely is, once in a while, it can be done, and it's extremely rare*
 P: [mm]
 ...
 P: yes, why do you say that once in a while, it can be done, but not
 D: no, so uh *in special types patients uh it is usually completely different kinds of diseases that we are talking about then*
 P: oh yes
 D: yes uh and *it's in an earlier stage you know, these are completely different kinds of lung diseases we're talking about*
 D: uh *a little earlier stage uh when we are talking about cancer, uh and lung cancer, then, then, then*
 P: oh yes
 D: a lung transplant . it's not feasible (...) (Ahus)

In excerpt 3.10, D conveys three warrants in support of his claim:

1. Lung transplants are usually infeasible
2. Lung transplants can be feasible for other diseases than cancer / Lung transplants are infeasible in the case of cancer
3. Lung transplants can be feasible in the early stage of a disease

It can be noted that while the second and third warrants are linked specifically to D, the first concerns lung transplants in general and does not seem to be clearly associated with a datum (see section 6.5.5).

Explanations can also be elicited in response to questions concerning how a datum explains a claim. In excerpt 3.2, when the patient asks what the datum (low oxygen saturation) means and whether it could be a sign of poor lung function (presumably in reference to the doctor's previous recommendation to do a lung X-ray), the doctor responds that he wants "to see that there isn't anything obvious in the lungs". This response can be interpreted as providing a missing link in a chained inference (see section 3.4.5).

A consistent pattern across analysed dialogues is that claims are supported *enthymematically* (Breitholtz 2020), i.e., by either data or warrants, rather than as argumentatively "complete" combinations thereof. For example, when D in excerpt 3.10 supports a claim with three warrants (e.g. that lung transplants can be feasible in the early stage of a disease), corresponding data (e.g. that the patient's disease has progressed far) are not verbalised, but can be understood as being presupposed. (For a discussion concerning modelling of enthymematic explanations, see section 4.5.5.)

With respect to dialogue modelling, the analysis indicates a choice whether explanatory information should be delivered (a) on demand, e.g. after the claim (as in excerpt 3.2), (b) before the explanandum has been presented (as in excerpts 3.4 and 3.5), or (c) concurrently with the explanandum/claim (as in excerpts 3.1, 3.6, and 3.7). There is also a choice between explaining claims with data or warrants (cf. section 4.5.5). Another modelling choice concerns whether the system expresses data with high numerical precision (as occasionally in the studied doctor–doctor dialogues), or with numerically less precise warrant triggers (as consistently in the studied doctor–patient dialogues). The path taken here is to assume that the system supports all of these strategies, and that a policy can be defined for choosing which strategy to use in a specific situation (see chapter 6).

An example of a rewritten *datum*-based explanation is given below:⁷

(3.1') s: I suggest switching medication, *since you have had your current medication for a relatively long time and have increased dosage and not had any effect.*

A rewritten *warrant*-based explanation is exemplified below:

⁷Rewritten excerpts are denoted with the original excerpt number followed by a prime.

- (3.10') U: do you think I can have these lungs replaced?
 S: no, I don't think that a lung transplant is feasible.
 U: why not?
 S: *Lung transplants are not feasible in the case of cancer.*
 U: mm
 S: *Also, lung transplants are not feasible in later stages of a disease.*

Note that the first warrant in excerpt 3.10 (lung transplants are usually infeasible) is excluded from the rewritten dialogue since base-rate explanations are not supported by the proposed dialogue model (see section 6.5.5).

3.4.3 Explanatory underspecification

Although linguistic triggers can help explainees identify potentially relevant warrants, a certain amount of explanatory underspecification (ambiguity) can be observed. One source of ambiguity concerns the potential existence of unexpressed intermediate reasoning steps. For example, in excerpt 3.2, D's use of warrant triggers (*lower oxygen percentage in the blood than one would expect*) still begs the question of what it is more precisely that low oxygen saturation may indicate, and that a chest X-ray can detect, as evidenced by P's follow-up questions (see section 3.4.5).

Ambiguities also concern whether implicated warrants are causal or statistical (or of some other type), as in the excerpt below:

- (3.11) D: It turns out the chest X-ray was completely normal and I couldn't see anything unusual in your throat or hear anything in your lungs that shouldn't be there. So I conclude that your symptoms aren't related to any illness in your respiratory system. (L&A, p. 251)

When D in excerpt 3.11 judges that P has no respiratory illness partly on the basis that the chest X-ray was normal, this is compatible with either a causal warrant (respiratory illnesses cause abnormalities that can be detected in a chest X-ray) or a statistical one (a normal chest X-ray correlates with absence of respiratory illness). In potential contrast with this, when D in excerpt 3.1 talks about having had a medication for over two months and increased dosage without "any effect", the use of causal language ("effect") may indicate a causal warrant such as "if a medication doesn't have the intended effect, one should try another medication". However, it is not unreasonable to conceive of a statistical warrant such as "in most cases where a medication hasn't worked after a long time and after having

increased dosage, switching medication is usually a good choice”.⁸

Furthermore, triggers can be underspecified with respect to the specific warrant content. Conveying oxygen saturation as lower than expected is compatible with a warrant that posits a *monotonically decreasing* relation between oxygen saturation and the probability of lung disease. (A relationship is monotonic if, as one quantity consistently increases or decreases, the other also changes consistently in one direction.) However, it is also compatible with a *non-monotonic* relationship, i.e. that too *high* oxygen saturation also indicates a higher risk of disease. Similarly, when multiple pieces of data are presented in support of a claim (such as having used a medication for a long time with a high dosage), potential interactions between data remain unstated. This kind of underspecification can potentially be understood as serving the purposes of relevance and brevity, i.e. only presenting information that is deemed relevant for the patient, and not providing more information than needed in the context (Grice 1975).

Potential dependencies between pieces of evidence constitute yet another source of ambiguity. When D in excerpt 3.10 gives three reasons for why a lung transplant is infeasible for P, those reasons intuitively seem to explain D’s assessment *independently* (as separate warrants). Specifically, the facts that lung transplants are infeasible in the case of cancer, as well as in a progressed stage of disease, may seem valid in their own right, irrespective of each other. However, this (potential) independence is not made explicit in the dialogue. In contrast, when D in excerpt 3.1 suggests switching medication since P has had a medication for over two months and increased dosage without any effect, there seems to be a dependency between pieces of evidence. Presumably, the long duration and increased dosage support switching medication only in the absence of an effect; if a desired effect had been obtained, the long duration and the increased dosage had not evidently pointed towards switching medication. (In other words, the three pieces of evidence might be associated with a single warrant.) Again, however, the (potential) *dependence* between pieces of evidence is not stated explicitly.

3.4.4 Turn-taking and complex explanantia

In most cases when interlocutors support inferences with multiple explanantia, the explanantia are delivered in a single turn with conjunctive content, as in excerpt 3.1: “since you have had over two months and have increased the dosage and not had any effect ...” In contrast, in excerpt 3.10 (repeated in part below), D delivers

⁸Intervention-related features are not uncommon in statistical (non-causal) ML. For example, the frequently cited German credit dataset (Hofmann 1994) predicts credit risk partly on the basis of previous credit decisions, while the statistical AI tool described in chapter 5 estimates the probability of successful outcome of surgery partly on the basis of whether the patient is previously operated.

additional warrants (italicised) after acknowledging feedback from P (bolded):

- P: yes, why do you say that once in a while, it can be done, but not
D: no, so uh in special types patients uh it is usually completely different kinds of diseases that we are talking about then
P: **oh yes**
D: yes uh *and it's in an earlier stage you know , these are completely different kinds of lung diseases we're talking about*
D: uh *a little earlier stage uh when we are talking about cancer, uh and lung cancer*, then, then, then
...

It is also worth noting that D's delivery of additional evidence is explicitly marked as a topic continuation with the discourse marker "*and (it's in an ...)*".

With respect to modelling, it is here assumed that an interlocutor who delivers complex explanantia has some means of designing a suitable turn-constructional unit (Sacks, Schegloff and Jefferson 1974) at any given moment in the dialogue, based on evidence available to the interlocutor (and the current state of the dialogue). For example, an interlocutor may have knowledge about three explanantia, but decide to initially only mention one of the three.

3.4.5 Chained inferences

As discussed in sections 3.4.1 and 3.4.2, interlocutors can interrogate and construct argument chains. In excerpt 3.5, the doctor presents evidence whose implications (forward expansions) are interrogated by the patient in two consecutive steps, thereby constructing a chained inference:

- Datum 1:** Bronchoscopy of P showed cellular changes
Claim 1 = Datum 2: P has a tumor (3.I)
Claim 2: P has cancer

An inference chain can also be revealed by unpacking or "zooming in" on an argument. When P in excerpt 3.2 asks what her low oxygen saturation means and whether it could be a sign of poor lung function, and D responds by stating that he wants to see that there "isn't anything obvious in the lungs", this can be understood as unpacking an argument into a chained inference and revealing an intermediate claim:

- Datum 1:** P has low oxygen saturation
Claim 1 = Datum 2: There is a non-negligible risk that P has a lung abnormality (3.II)
Claim 2: Doing a chest X-ray on P is an adequate action

This analysis raises the question: Why does D initially explain his intent to perform a lung X-ray with reference to the *inferentially distant* datum (corresponding to datum 1 in schema 3.II), rather than with the datum that inferentially supports the claim in more *proximately* (i.e. datum 2 in schema 3.II)? One potential reason might be an intent to not induce anxiety (as discussed in section 3.4.8). Since it is difficult to explain the proximate inference without mentioning the risk of a potential lung abnormality, D's initial explanation skips the intermediate reasoning step and focuses on the inferentially more distant datum, which is presumably deemed less anxiety-inducing. Another reason might be that D selects the empirically most grounded explanation. The low oxygen saturation has been measured objectively, while the non-negligible risk of a lung abnormality is inferred inductively on the basis of the objective measurement, and is therefore less certain.

Chained inferences relate to the semantics of target labels in ML. Many ML datasets in the medical domain are trained to predict outcomes of future measurements, rather than an appropriate action or intervention. For example, the frequently used UCI Heart Disease dataset (Janosi et al. 1989) has as ground truth the number of major blood vessels detected with coronary angiography (which, in turn, can indicate heart disease). In other words, a model trained on the Heart Disease dataset does not predict whether a certain heart examination or treatment is appropriate, but only what a coronary angiography might reveal. Based on such a prediction, an adequate action can be recommended via an additional inference step. Similarly, a dialogue system serving the role of the doctor in excerpt 3.2 can be assumed to use a model that estimates the probability that a chest X-ray would reveal an abnormality; the system then decides whether to recommend chest X-ray based on the model's estimated probability and, e.g., some pre-configured threshold. (The reasoning behind such a design would be that a chest X-ray should only be taken if there is a non-negligible probability that it would reveal something abnormal.) Based on this assumption, excerpt 3.2 is rewritten as follows:

- (3.2') s: I recommend performing a chest X-ray.
u: why is that?
s: The level of oxygen in your blood is relatively low.
u: oh? how high is it?
s: It's at ninety-three percent.
u: could a too low oxygen uptake indicate that lung function is poor?
s: Yes, low oxygen saturation is statistically associated with a high risk of lung abnormality.

In excerpt 3.2', S' initial explanation corresponds to datum 1 in schema 3.II,

and U's final question is interpreted by the system as concerning a warrant for the same inference step. It is here assumed that when the system explains a claim resting on a chained inference, the system has some means of determining which inference step to focus on. In excerpt 3.2', the system focuses on the inferentially distant datum, while in the (more hypothetical) excerpt below, it focuses on the inferentially proximate datum:

- (3.2'') s: I recommend performing a chest X-ray.
u: why is that?
s: I suspect that an X-ray might reveal something abnormal.
u: why do you think so?
s: The level of oxygen in your blood is relatively low.
...

3.4.6 Counter-argumentation

Occasionally, interlocutors elicit evidence against (own and other's) claims, as in excerpt excerpt 3.4 (repeated for convenience below):

- D: Also, I want to inform you that your blood tests show your blood sugar is perfectly normal, and your cholesterol levels are good too.
P: So I'm fairly healthy for my age?
D: Yes, I'd say you are. *Except for your high blood pressure, I mean your blood pressure condition, you have that (...)*

In excerpt 3.4, D's mention of P's high blood pressure opposes D's own claim that P is relatively healthy. Arguments against the other interlocutor's claim can also be observed, as in the following excerpt, where R argues that the patient should be given fiberoptic intubation orally, while A wants to do it nasally:

- (3.12) R: Let's just try an oral for once, see if we can do it
 ...
 A: When he's asleep I'll tell you why I didn't do an oral
 ...
 A: I just think with orals you really have to sedate him a lot more...
 R: See I don't. I've done a lot of them, you don't have to sedate them at all
 A: (..) that big airway...
 R: Absolutely.
 A: But you should be able to do them both
 R: No, I just think, my choice with a guy like this with very small nares, and he's this large, I rather have a bigger tube, and to do that you have to go oral
 (Johannesen, p. 41)

With respect to modelling, the system is here assumed to primarily function as a recommender. Therefore, excerpt 3.12 is rewritten to two separate dialogues, where, in each excerpt, the system argues for one of the claims:

- (3.12'o) S: I suggest oral intubation.
 U: Why?
 S: The patient needs a big tube.
 U: Why do you think so?
 S: The patient has small nares and a big body.
- (3.12'n) S: I suggest nasal intubation.
 U: Does the patient's big airway point towards going oral?
 S: Yes, the patient's big airway is a reason for choosing oral intubation.

Note that in excerpt 3.12, A's indirect acknowledgement that a big airway supports the oral route indicates a potential value in enabling the system to confirm evidence that contradicts its prediction (cf. T. Miller 2023), as exemplified in the rewritten excerpt 3.12'n. Also note that in the rewritten excerpt, the system's confirmation of counter-evidence is triggered by a user question, in contrast to the original excerpt 3.12, where the trigger is unclear. Furthermore, arguments focusing on whether a particular type of intubation is generally favourable (since it requires less sedation) are excluded (see section 6.5.5).

3.4.7 Hedging and epistemic stance

Interlocutors frequently hedge their claims with epistemic markers such as “no, I *suppose that*, unfortunately, it isn’t (feasible)” (excerpt 3.3), “Yes, *I’d say* you are” (excerpt 3.4), and “*Have a feeling* it’s [hematocrit measurement] still gonna be low” (excerpt 3.6). With respect to modelling, it is here assumed that such forms of hedging reflect degree of confidence, and that information about the model’s predictive confidence (certainty) is available to the system.

3.4.8 Answer unavailability

Occasionally, interlocutors signal answer unavailability. In excerpt 3.2, D uses the word “unspecific” twice to indicate uncertainty. The first time, the uncertainty concerns the expected oxygen saturation (“it’s at ninety-three percent, and one would expect between ninety-five yes, I mean it’s a bit unspecific”). The second time, it concerns potential causes of the patient’s low oxygen levels (“in itself, it’s terribly unspecific”). Various cues in this particular dialogue indicate that D evades questions as a means to not worry the patient. For example, when D wants “to see that there isn’t anything obvious in the lungs”, the wording seems to purposely avoid mentioning the kind of lung abnormality that the low oxygen saturation might indicate. Furthermore, later in the dialogue, the doctor says that chest X-ray is taken “as a routine ... of almost everyone” and “since you’re already here, then we’ll check that as well”. In the context of the present chapter, such anxiety-reducing strategies are deemed out of scope, and or not modeled. However, as will be discussed in section 3.4.9 below, it can be appropriate for a dialogue system to signal answer unavailability for other reasons.

3.4.9 Ontic explananda

Even if this thesis focuses on epistemic rather than ontic explananda, the present analysis shows that the two type of explananda can be tightly linked to each other. In excerpt 3.2 (repeated for convenience below), an epistemic explanation (concerning why the doctor wants to X-ray the patient’s lungs) leads to an ontic explanation (concerning the causes for the patient’s low oxygen saturation):

- D: I just feel like taking an X-ray of your lungs as well, on the way up,
P: why is that?
D: no, just, because I looked at the oxygen percentage in your blood,
which is a bit lower than one would expect,
P: oh? how high is it?

- D: uh it's at ninety-three percent, and one would expect between ninety-five yes, I mean it's a bit unspecific,
- ...
- P: *what does it mean to have too low oxygen uptake then? what could it be a sign of? that lung function is poor?*
- D: uh in itself, it's terribly unspecific, I just want to see that there isn't anything obvious in the lungs (...)

The emphasised question from P can be interpreted as concerning the physiological cause of the low oxygen saturation, at least in a Peircean interpretation of “sign of” (Peirce 1992). As dialogue modelling is concerned, this suggests that the system should be able to respond appropriately to users' follow-up explanation questions to explanatory datum assertions made by the system, even when such information cannot be derived from the statistical model. To favour transparency, it is in such cases deemed appropriate for the system to signal answer unavailability:⁹

- (3.2'') S: I recommend performing a chest X-ray.
 U: why is that?
 S: The level of oxygen in your blood is relatively low.
 U: oh? how high is it?
 S: It's at ninety-three percent.
 U: could a too low oxygen uptake indicate that lung function is poor?
 S: Yes, low levels of oxygen in the blood can indicate a lung abnormality.
 U: *what might be the cause for my low oxygen uptake?*
 S: *I don't have any information regarding the causes for the low oxygen levels.*

3.4.10 Feedback and grounding

Interlocutors use various forms of feedback such as “oh yes” and “hm?”, as in excerpt 3.10 (repeated for convenience below):

- P: [do you think] I can have these lungs replaced?
 D: [yes, that would have been] really nice, that
 D: *hm ?*

⁹In the rewritten excerpt, U's emphasised question is purposely phrased so as to interrogate the cause of the datum, rather than how the datum explains the claim.

- P: it's not feasible
 D: no, I suppose that, unfortunately, it isn't
 D: it rarely is , once in a while, it can be done, and it's extremely rare
 P: [*mm*]
 ...
 P: *yes*, why do you say that once in a while, it can be done, but not
 D: no, so uh in special types patients uh it is usually completely different kinds of diseases that we are talking about then
 P: *oh yes*
 D: yes uh and it's in an earlier stage you know , these are completely different kinds of lung diseases we're talking about
 D: uh a little earlier stage uh when we are talking about cancer, uh and lung cancer, then, then, then
 P: *oh yes*
 D: a lung transplant . it's not feasible (...)

In excerpt 3.10, D reacts to P's acknowledgements by providing (additional) support for his claim or, in the last instance, by repeating his claim. With respect to dialogue modelling, an acknowledgement from the user is treated as a signal to continue delivering more information, if such is available (see section 3.4.4); if not, the acknowledgement is silently ignored.

3.4.11 Anaphora

Anaphoric references to explananda, such as “why is *that?*”, are frequent in the corpus. During the process of distillation, anaphoric references are assumed to be contextually resolved by the NLU component.

3.4.12 Methodological elaborations

One of the main challenges when rewriting non-AI-assisted human–human dialogues to human–AI dialogues is that the analyst does not have access to the human interlocutor's “model” (mind). This makes it difficult to make confident analogies between human judgements and AI predictions. The lack of access concerns both the *types of evidence* that the human interlocutors might potentially consider when making their judgements (corresponding to ML features), the way in which the human interlocutors *combine* or *weigh* pieces of evidence into judgements (corresponding to the type of ML model, e.g. linear model or decision tree), and the range of potential judgements available to interlocutors (corresponding to target labels or sets of models). For example, when a doctor explains a

recommendation with reference to the low oxygen saturation (single feature), and no other pieces of evidence are mentioned in the dialogue, the analyst cannot determine if the doctor also (consciously or unconsciously) considered other pieces of evidence (features), such as the patient's age, medical history, symptoms, etc., and if so, how the features were combined or weighed into the judgement. Furthermore, one cannot know if the doctor also considered other actions or recommendations.

The approach taken here has been to assume that interlocutors consider *at least* the pieces of evidence that they refer to when explaining their judgements, but potentially additional pieces of evidence as well. In line with Grice's maxim of quantity (Grice 1975), it is also assumed that evidence considered by interlocutors, but not mentioned in explanations, have limited or no impact on their judgements. For example, if the doctor also considered the patient's heart rate when assessing whether a chest X-ray should be performed, then this factor is assumed to not have affected the judgement significantly, since if it had, this would have been mentioned.

It is also assumed that all judgements are binary, e.g. whether or not to perform a chest X-ray. This assumption is evidently oversimplified, since in real medical situations, the range of potential judgements that might be relevant in a given situation is very large. However, since the analyst cannot confidently know what that range might be, this aspect of the interlocutor's reasoning process is very difficult to model.

3.5 Discussion and conclusions

Based on a method for distilling human-computer dialogues from human-human dialogues in medical settings, this chapter has revealed a diverse range of dialogue phenomena and strategies which can inform the design and modelling of CXAI systems. On a general level of discourse dynamics, the observed strategies highlight the co-constructive nature of explanations (T. Miller 2019; Rohlfing et al. 2021; Finke et al. 2022; Buschmeier et al. 2025), where the process of communicating the epistemic basis of a complex judgement rarely occurs in the form of a one-shot delivery from explainer to explainee. Specifically, when the explanans is complex and consists of multiple pieces of evidence or circumstances, interlocutors jointly coordinate the process of unpacking the explanans, using feedback, backchannels and continuation markers. Furthermore, explanations often give rise to requests for additional explanations, where the explanans for the initial explanandum, or the relation between the explanans and the initial explanandum, itself becomes a new explanandum. The process of expressing and interpreting explananda is also

co-constructive, in that explainees often refer to previous assertions deictically, requiring the explainer to contextually infer what it is that should be explained.

From a rhetorical and argumentative perspective, the analysis reveals that explanations in the analysed data tend to be delivered enthymematically, i.e., as arguments where some premises are implicit and assumed to be identified by the explainee using context, prior knowledge and linguistic triggers. Occasionally, judgements are supported by a chain of inferences. Furthermore, interlocutors occasionally attack each other's arguments by providing counter-evidence and affirming each other's attacks.

Arguably, insights derived from studies of human–human explanatory dialogues can be potentially informative in relation to design and modelling of human–AI interaction, even if the judgements being explained in the studied human–human interactions do not involve AI. Observed strategies in human dialogues can be used as desiderata for conversationally explainable AI systems, by identifying specific capabilities that such a system needs to have in order to act as explainer in analogous human–AI interactions (see chapter 6). The lack of AI involvement can even be regarded as a benefit, in that interlocutors who explain their own judgements can be expected to address the epistemic bases for such judgements more “naturally” than when they explain predictions on behalf of an AI system whose mode of reasoning they do not necessarily understand.

Nevertheless, one limitation associated with using human–human dialogues to inform CXAI is the lack of access to ground truth epistemic bases. The only information available to the analyst concerning the actual reasoning behind the judgements expressed in analysed dialogues is the linguistic data. This situation can be conceived as problematic in light of previous findings concerning the unreliability of human explanations for their own judgements and behaviours (Nisbett and Wilson 1977). In this light, studies of human–human dialogical explanations can say something about how humans use language to interrogate and explain judgements in interaction, but not about how their statements relate to the actual causes or reasons behind those judgements. This limitation has at least two potential disadvantages. First, the prospect of informing design and development of CXAI systems by studying human–human dialogues hinges, at least to some extent, on an assumed potential analogy between how interlocutors reason and how ML models learn from data and make their predictions. However, given the lack of access to interlocutors' actual reasoning, it is unclear to what extent this potential analogy holds. Even if observed human explanations are potentially compatible with statistical inferences, as discussed in section 3.4.3, other interpretations are conceivable. For example, an interlocutor might reason causally, without making this explicit in dialogue. Second, when striving for CXAI systems that not only converse naturally and in a human-like manner,

and give plausibly sounding explanations for their predictions, but that also explain themselves in ways that can be *trusted*, it is important to understand the relation between reasoning and language use. However, this relation cannot be studied by merely observing language use. To address these limitations, chapter 4 experimentally collects human–human explanatory dialogues revolving around AI predictions in a setting where the explainer, but not the explainee, has access to information about the AI’s reasoning. This makes it possible to study how humans address the epistemic basis of judgements in a context where both the explainer and analyst have access to the same information about the reasoning behind the judgements.

It should also be acknowledged that the insights and research challenges derived in the present chapter pertain to a relatively small sample of dialogues in a single domain (healthcare). This limitation is to some extent mitigated in chapter 4, which investigates explanatory dialogues in another domain (personality assessment based on music preferences). Overall limitations associated with the employed methodology will be further discussed in section 7.2.

CHAPTER 4

Lay explanations in dialogues involving AI-assisted personality assessments

A limitation of the empirical study in chapter 3 was that the lack of information concerning ground truth epistemic bases for judgements impeded analyses of how the stated explanantia relate to the actual epistemic underpinnings of the judgements. This made it difficult to assess potential analogies between interlocutors' reasoning and ML inferences, and to study conditions under which explanations can be characterised as reliable. In order to address these limitations, the present chapter collects human–human explanatory dialogues in a setting where the explainer and analyst have access to the same epistemic information.

Specifically, the chapter collects and analyses lay explanations in web-based human chat conversations involving AI-assisted personality assessments. Based on how participants rate excerpts of music, an AI estimates their personality along five dimensions (extraversion, openness, etc.) After having rated the music, participants are paired up and chat to each other about the AI's outputs.

The main aim of the chapter is to investigate how non-experts conversationally address the epistemic basis of AI predictions, given information presented to one of the interlocutors in a primarily visual tool. By designing the communicative situation so that only one of interlocutors (the explainer/operator) has access to the AI's output for the other interlocutor (the explainee/respondent), the interlocutors are incentivised to interpret and linguistically coordinate their comprehensions of the AI's predictions. Unlike naturally occurring explanatory interactions without AI involvement, the setting makes it possible to study how interlocutors interrogate and address the epistemic basis of judgements in a context where the analyst has access to the same epistemic information as the explainer, and where the judgements are explicitly derived from AI models. Specifically, the setup enables studying whether stated explanations are faithful with respect to the actual workings of the AI, to what extent explainers refer to the AI versus other sources of evidence, and how interlocutors handle explanatory needs that cannot straightforwardly be addressed merely on the basis of information about the AI.

Since AI predictions and their epistemic bases concern the participants as such (rather than, e.g., fictional personas), no role-playing is required, which benefits ecological validity. Based on an assumption that many humans are interested in

music and their own personality, it is hoped that participants will spontaneously be motivated to understand the bases of the AI's predictions, and therefore to address such issues in a chat conversation.

4.1 Related work

Previous work has collected and/or analysed explanatory dialogues revolving around ML predictions in setups where the explainer is a dialogue system (Kuźba and Biecek 2020) or a researcher acting as the system (Diana C Hernandez-Bocanegra and Ziegler 2021). In contrast to previous work, the present chapter collects explanatory dialogues revolving around model predictions, with human participants/informants in the roles of both explainee and explainer. This makes it possible to study human dialogue strategies in relation to ML predictions while at the same time avoiding risks associated with the use of confederates (Kuhlen and Brennan 2013). The present chapter also extends previous work by identifying dialogue capabilities manifested in human–human interactions which might be useful for CXAI systems to possess.

4.2 Experimental setup

The data collection takes the form of a browser-based experiment where participants first listen to 30-second excerpts of 10 music tracks and rate them on a 4-point hedonic scale (dislike very much, dislike slightly, like slightly, like very much). When a participant has rated all tracks (and their ratings contain variation), they can proceed to the second part of the experiment. In the second part, participants are paired up with each other and are randomly assigned the role of either respondent or operator. They then chat with each other using an interface (see figure 4.1). Operators are instructed to explain the respondent's test results, while respondents are instructed to ensure that they receive their test results and to try to understand what they are based on. Operators, but not respondents, are given access to prediction results (estimated personality traits), information about the statistical model, definitions of personality traits, local and global feature contribution plots (see figures 4.2 and 4.3), and feature values (plots of the respondent's music preferences; see figure 4.4).

Data was collected through a series of 5 trials; the last three of these trials included an additional third part where participants were paired up a second time after completing their first chat, but this time in opposite roles (potentially with another participant than in the first chat). In the last two trials, demographic

information (age, gender and educational level) was collected on a voluntary basis after the two chat interactions.

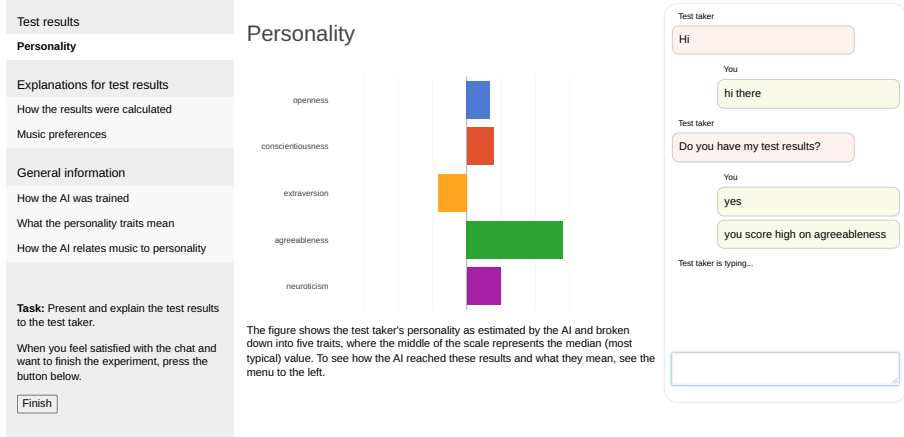


Figure 4.1: Screenshot of operator’s main view during chat, with hypothetical prediction and chat messages. In the personality prediction plot, the X axis reflects log odds; bar lengths are proportional to the total estimated log odds for the respective traits ($\sum_i X_i \beta_{ti}$). Respondents only see a chat window (similar to right-most part of operator’s view).

Since participants are paired up with each other, known issues of bias when using confederates (Kuhlen and Brennan 2013) are avoided, enabling an open-ended investigation. When participants could not be paired up with each other, they were paired up with the experiment leader (the author). Data collected in this manner was excluded.

For each big-five personality trait (John and Srivastava 1999), a logistic regression model is trained to predict whether an individual has the trait in question (e.g. extraverted) based on the individual’s music preferences and psychometric test results as ground truth. The dataset consists of music listening data and psychometric test results for 1000 users of the music website Last.fm (Melchiorre and Schedl 2020a).¹ Specifically, for each trait, the training set is divided into two equally sized partitions: positive instances (trait value above median) and negative instances (remaining instances). In other words, the models are trained to estimate whether an individual belongs to the upper or lower half of the trait spectrum. For interpretability purposes, no intercept/bias terms are used.

¹The dataset is available in a public repository (Melchiorre and Schedl 2020b).

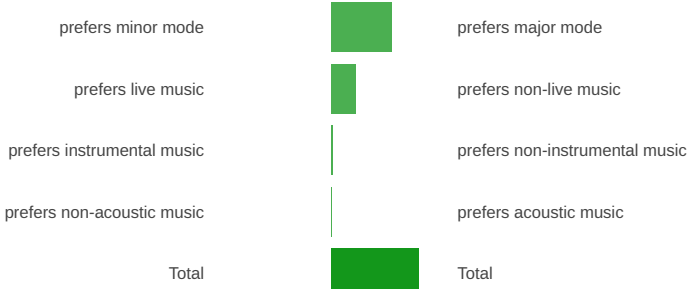


Figure 4.2: Example of a local feature contribution plot for agreeableness, in this case with only positive contributions. These plots are showed to operators under the section “How the results were calculated”, with the following caption: *The figures above show how the AI sums up scores for different music preferences to scores for personality traits.* Contributions are sorted by decreasing importance with respect to the specific prediction. The X axis reflects log odds; for features, bar lengths are proportional to contributions ($X_i\beta_{ti}$), while for the “Total” bar, the length is proportional to the total estimated log odds ($\sum_i X_i\beta_{ti}$), as in fig. 4.1.

Music preferences are encoded numerically as standardised aggregated feature values for audio properties (energy, loudness etc.):

$$X = Z_S(\tilde{X})$$

where \tilde{X} is a vector of raw aggregated (non-standardised) feature values. The function Z_S standardises a vector with respect to a sample S :

$$Z_S(\tilde{X}) = \frac{\tilde{X} - \mu_S}{\sigma_S}$$

where μ_S and σ_S contain means and standard deviations for sample S . For instances in the dataset, \tilde{X} contains mean values across each track that the individual has listened to. For example, if an instance in the dataset has $\tilde{X}_i = -10$ for loudness, this means that the tracks listened to by the individual have a mean loudness value of -10 . During model development, standardisation is performed with respect to the entire dataset.

Number of features range from 2 to 5 depending on trait (see figure 4.3) and were selected manually as a trade-off between predictive performance and sparsity.

During inference, feature values are aggregated using weighted averaging across the rated tracks:

$$\tilde{X}_i = \frac{1}{\sum_{k=1}^n r_k} \cdot \sum_{j=1}^n r_j a_{ij}$$

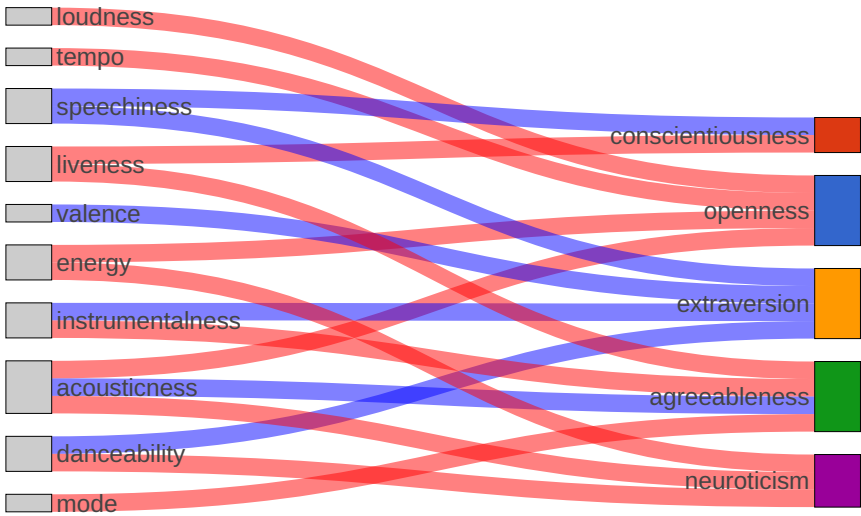


Figure 4.3: Global feature contribution plot, showed to operators under the section “How the AI relates music to personality”, with the following caption: *The figure shows correlations between music preferences and personality traits, as learned by the AI. Blue links correspond to positive correlations, red links to negative correlations. For example, a red link from loudness to openness means that a preference for silent music correlates with openness, and conversely that a preference for loud music correlates with non-openness. (Higher valence means a stronger preference for happy/cheerful/euphoric music. Higher mode means a stronger preference for music in minor mode.) Blue links reflect positive coefficients ($\beta_{ti} > 0$) while red links reflect negative coefficients ($\beta_{ti} < 0$). Note that coefficient magnitudes are not visualised, based on the assumption that such level of detail would make the plot more difficult to interpret.*

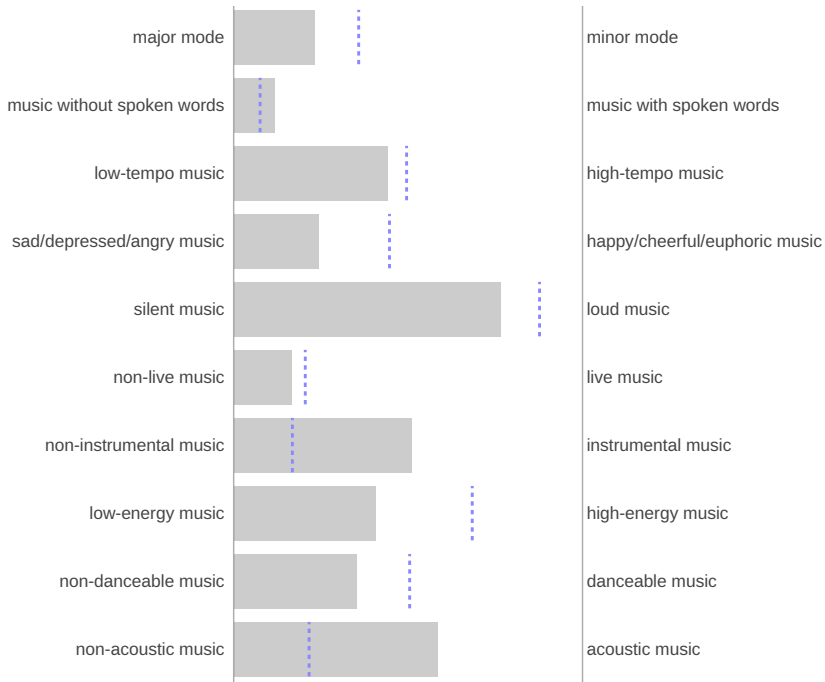


Figure 4.4: Plot of feature values for the respondent in relation to mean levels for the rated tracks, showed to operators under the section “Music preferences”, with the following caption: *The figure shows the test taker’s music preferences compared to average levels for the rated tracks (dotted blue lines), based on how the test taker rated the tracks in this experiment. Audio data for each track comes from Spotify. The X axis spans from 0 to 1 and reflects raw feature values scaled to feature-specific minimum and maximum values; in other words, a value of 0 represents the minimum feature value, while a value of 1 represents the maximum feature value. Bar lengths reflect the respondent’s aggregated feature values (\bar{X}_i) while dotted blue lines reflect mean values for the rated tracks.*

where r_j is the respondent's rating for track j , and a_{ij} is the raw audio property value of feature i and track j . For participants, standardisation is performed with respect to the 10 rated tracks. Traits are predicted by estimating the probability that a respondent belongs to the upper half of the trait spectrum:

$$P(Y_t = 1 | X) = g^{-1} \left(\sum_i X_i \beta_{ti} \right)$$

where Y_t is the value of trait t and β_{ti} are the regression coefficients. The inverse link function g^{-1} for logistic regression is defined canonically:

$$g^{-1}(\eta) = \frac{1}{1 + \exp(-\eta)}$$

The tracks to be rated by participants were selected so as to maximise feature variance and thereby create ideal conditions for predictive performance. Specifically, a pool of candidate tracks was first obtained using Spotify API's recommendations method (Spotify 2022). (Oral genres such as comedy and writing were excluded.) 100 random subsets, each containing 10 tracks, were then created. Finally, the subset with maximum variance (total standard deviation) across normalised feature values was selected.

It should be noted that due to the interpretability of the models, and the way in which explanatory information is obtained and visualised, the explanations in the tool accurately reflect how the models make predictions. For example, if a local feature contribution plot presents a high estimated agreeableness as mainly being composed of a preference for music in major mode (see figure 4.2), this information is inherently faithful with respect to the actual workings of the model. (The specific methods for visualising explanatory information are elaborated in the captions of figs. 4.1 to 4.4.)

Since English proficiency levels are generally very high in the local population, and since potential participants were informed in the recruitment material that they would interact with each other in English and could choose to opt out on this basis, comprehension checks were not deemed relevant.

4.3 Recruitment of participants

Since the experiment focuses on lay explanations, participants were recruited from a general audience where individuals cannot be expected to have expert knowledge concerning the relation between music preferences and personality or how predictions from statistical models can be interpreted and explained.

Specifically, in initial trials, recruitment was performed using convenience sampling of colleagues at the Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, and students attending a course on AI in society at the university; in subsequent trials, participants from the general public were recruited via information at the university’s website, posters and flyers at the university campus, and marketing in social media. In the recruitment material, potential participants were invited to participate in an experiment where an AI assesses one’s personality based on one’s music preferences, and were informed that the assessment results would be communicated in English between participants. No monetary or other reward was offered to participants.

4.4 Method of analysis

Collected dialogues were distilled using the method described in section 3.3. The dialogue system acting as substitute for the operator is assumed to use the same set of statistical models as in the tool used by operators. As in chapter 3, the dialogue distillation was performed by the author, partly on the basis of feedback from the author’s main supervisor who reviewed in-progress distillation results.

4.5 Results

74 individuals participated in the experiment across 5 trials between June 2022 and September 2024, resulting in a corpus of 35 dialogues (referred to below as “D 1” etc.) encompassing 779 utterances (disregarding dialogues where a participant was paired up with the experiment leader). Trial-specific information and descriptive statistics for collected data are provided in table 4.1. Collected information about participant demographics (gender, age and education level) are summarised in table 4.2.

The entire collected dataset is available in the Dialogues on Music, Personality and AI repository (Berman 2025b). Automated tests demonstrating that the rewritten dialogues are implementable using the dialogue model outlined in chapter 6 are available in the repository for the dialogue manager (Berman 2025a).

9 (26%) of the collected dialogues addressed epistemic bases of AI predictions. Other common topics, all of which were excluded from the analysis, include the meaning of psychological terms (openness, extraversion etc.), coordination of the situation as such (e.g. questions concerning whether to proceed the next step of the experiment), as well as discussions concerning potential agreement or disagreement with AI predictions.

Table 4.1: Trial-specific information (recruitment channel, date of data collection, and potential number of chats per participant) and descriptive statistics for the collected data. “Excluded” refers to the number of dialogues where a participant was paired up with the experiment leader; these are not included in “Dialogues”, which refers to the number of collected dialogues between participants. “In scope” refers to the number of dialogues that were revealed during analysis to address epistemic bases of AI predictions.

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	<i>Total</i>
Channel	Colleagues	Colleagues	Students	Public	Public	
Date	Jun 2022	Jun 2022	Apr 2024	Sep 2024	Oct 2024	
Chats/part.	1	1	2	2	2	
Participants	3	4	15	46	6	<i>74</i>
Utterances	113	114	190	284	78	<i>779</i>
Dialogues	1	2	12	18	2	<i>35</i>
Excluded	1	0	0	5	2	<i>8</i>
In scope	1	1	4	2	1	<i>9</i>

A distillation of the dialogue excerpts of interest revealed various phenomena and challenges concerning design and modelling of human–AI explanatory dialogue (henceforth referred to simply as *dialogue modelling*), as elaborated in the subsections below.

4.5.1 Types of explananda

Observed epistemic explanations occasionally target numerically conveyed predictions, as in the excerpt below (O=operator, R=respondent):²

- (4.1) O: You scored 5 on Openness. You scored 0,5 on Conscientiousness. ...
 ...
 R: Why is it a 5 on Openness ... (D 2)

However, in most cases, predictions are interpreted and framed by the operator when conveyed to respondents, as in the excerpt below:

- (4.2) O: You have been rated *the highest in openness*
 R: Oh wow, why? :) (D 15)

²Emphases in cited corpus excerpts have been added by the author.

Table 4.2: Participant demographics for the trials where such information was collected.

	Trial 4	Trial 5	Total
Gender			
Female	7	2	9
Male	9	2	11
Other	1	0	1
Age			
18–24 years old	1	0	1
25–34 years old	3	0	3
35–44 years old	2	1	3
45–54 years old	6	2	8
55–64 years old	3	1	4
65–74 years old	1	0	1
75 years old or older	1	0	1
Education level			
No schooling completed	0	0	0
Completed high school / gymnasium	0	0	0
Completed university degree	17	4	21

Notably, in excerpt 4.2, O’s framing of the prediction involves a subjective evaluation of how the graphically presented information can/should be interpreted and contextualised in terms of how low/high a particular value is. A similar behaviour can be observed in the excerpt below, where O both frames the prediction evaluatively and elaborates the value numerically in more objective terms:

- (4.3) R: What kind of information does the test give you?
 ...
 O: apparently, you are *very open*
 O: *almost 5* (out of -5 to 5 where 0 is the median)
 R: It’s interesting, I wonder what song would give this trait
 (D 13)

While most explananda target *polar* evaluations of predictions (i.e. that an estimated value is deemed low or high), in one case, the operator frames a prediction as *neutral* (ambivalent):

- (4.4) O: on extroversion, you scored *pretty close to the median*
 R: Do you know the link with music features? (D 13)

With respect to dialogue modelling, it is here assumed that the dialogue system frames predictions polarly in relation to a median music listener, similar to how predictions are visualised in the tool used by operators (see figure 4.1). For example, if the system expresses that the user is “open”, this reflects that the user’s estimated score on openness is higher than for a median music listener.

4.5.2 Explanation triggers and query types

In many cases where operators provide explanations, this is triggered by explicit queries by respondents. Among explanation requests, wh-questions are the most common. Examples include “what do you base this conclusion on”, “Why do you think I’m a very agreeable [sic] person?” and “Among my answers, which ones tilted me the most towards being non-agreeable and slightly open?”

In some cases, yes-no questions are used. In one such case, the question pragmatically serves as a confirmation request: “But the result could be because of a preference of music with lower tempo / am I correct?” In another case (excerpt 4.4), the question “Do you know the link with music features?” pragmatically serves as a wh-question concerning a link between the respondent’s ratings and the AI’s judgement.

In two instances, literal assertions pragmatically serve as queries. In one such instance, R’s utterance literally conveys an interrogative stance towards the embedded clause:

- (4.5) R: *I wonder if music influences the personality or if it’s only the other way*
O: yeah I was thinking that too (D 13)

In another instance, R’s syntactically assertive utterance can be pragmatically understood as a confirmation question:

- (4.6) O: the AI calculates the results based on a statistical model for each personality trait
...
R: hm ok, so *it takes into account many other people’s statistics then*
O: a 1000 users apparently (D 19)

In excerpt 4.6, O’s response seems to felicitously serve the broader information need implicated by R’s indirect confirmation question.

With respect to dialogue modelling, identification of intended query type and content is here assumed to be performed by the NLU component. How query

types relate to the system's choice of response strategy is discussed in section 4.5.3 below.

4.5.3 Types of explanantia

In response to queries concerning the basis for the AI's assessments, varying response strategies are used. One observed strategy is to *provide the source of evidence*, without detailing how specific types of information relate to the assessment:

- (4.7) R: what do you base this conclusion on
O: This conclusion is based on *the score from your ratings of the music you listened to* (D 2)

A related strategy is to *mention the factors* on which the AI's assessment is made, without detailing how the factors contribute to the assessment:

- (4.8) O: You scored -2,3 on Extraversion, which tells me that you are likely more introverted.
...
O: The Extraversion score *was based on Dancable/Non-Dancable music, Happy/Sad music, Intrumental/Non-instrumental music and music with/without spoken words.* (D 2)

In another case, the operator produces an *algorithmic narrative* conveying how the assessment was calculated. The narrative straightforwardly transforms the feature contribution plot into natural language:

- (4.9) R: Why is it a 5 on Openness, based on what variables ...
O: *The 5 on Openness comes from the total score of four Preferences. Loud music is -5 and silent music is 5—you scored 4.5. Other preferences were acoustic music at -5 and non-acoustic on 5, where you scored -1. The high-energy music preference was -5 and low-energy music was 5—you scored 3,8. High-tempo preference was -5 and low-tempo music was 5—you scored 0,8. The total of these 4 preferences were a 5 on Openness* (D 2)

In other cases, the operator explains the AI's assessment *argumentatively*, i.e. as a claim supported by evidence:

- (4.10) R: Can I get my test results? What are the scores and their interpretation?
 ...
 O: you have scored low on “openness” and “neuroticism”
 O: slightly higher on “extraversion”
 ...
 R: I really want to know what these results are based on...why am I so low on openness? kind of disagree with that
 O: “openness to experience”: *the score was calculated on the basis that you don't prefer high-energy, loud music*
 ...
 O: “Conscientiousness” *was based on the fact that you prefer non-live music and music with spoken words*
 O: “Extraversion” *is based on the fact that you prefer music with spoken words while you disprefer sad/depressed/angry music, non-instrument music, non-danceable music* (D 19)

Argumentative strategies are observed both for model-grounded explanations, as in excerpt 4.10 above, and for explanations grounded in the operator’s own knowledge, as in the excerpt below, where O explains the negative score for neuroticism with reference to R’s preference for death metal:

- (4.11) O: Your neuroticism was -1.7
 ...
 O: Do you *not enjoy death metal*
 R: not particularly
 R: but slightly
 O: Fair enough
 O: Explains the slightly negative score (D 15)

In another instance, O speculates that not liking a particular track supports being either neurotic or non-neurotic:

- (4.12) O: You have high on openness, and very low on neuroticism. neutral on the rest
 R: oh, low on neuroticism, but I pressed so fast on the growling-dislike-button!
 ...
 O: *Maybe if you were neurotic you would not like the growling, or you could. I can see reasons in both directions* (D 34)

Observed argumentative strategies typically focus on *data*, i.e. specific grounds

supporting the prediction. In most cases, data are conveyed polarly, e.g. “you don’t prefer high-energy”, “you showed a preference for acoustic music”. (In technical terms, this corresponds to characterising a feature level as either low or high.) However, in one situation where the prediction itself is neutral, data is also conveyed neutrally (i.e. as neither low nor high):

- (4.13) O: on extroversion, you scored pretty close to the median
R: Do you know the link with music features?
O: in the explanations chart there is a strong relationship between extroversion and *dancability and liveliness*
O: but if I’m reading this correctly, *your preferences didn’t indicate strongly one way or another about those features*
(D 13)

In another instance, a neutral claim is supported by contradictory evidence:

- (4.14) O: you have scored low on “openness” and “neuroticism”
O: slightly higher on “extraversion”
...
R: I really want to know what these results are based on ...
...
O: “Extraversion” is based on the fact that *you prefer music with spoken words while you disprefer sad/depressed/angry music, non-instrument music, non-danceable music*
(D 19)

In excerpt 4.14, O explains a neutral estimate of extraversion using a contrastive while-clause whose main clause (“you prefer music with spoken words”) reflects evidence for extraversion while the subordinate clause (“you disprefer sad/depressed/angry music, non-instrument music, non-danceable music”) reflects evidence for introversion.³

While observed arguments typically highlight data, in the following exchange, O also conveys *warrants* (general rules or patterns, in this case relations between music preferences and personality):

³A similar behaviour can be observed later in the same dialogue in excerpt 4.24, although in that case, the prediction is quite far from the median.

- (4.15) O: apparently, you are very open
 O: almost 5 (out of -5 to 5 where 0 is the median)
 R: It's interesting, I wonder what song would give this trait
 O: well I actually can tell you something about that I think
 O: not which song in particular, but how openness relates to
 features of the music
 R: Oh great I'm interested
 O: for example, *openness apparently has a strong relationship
 with acousticness*
 O: and I think this is saying you showed a preference for
 acoustic music
 R: Okay, yes I think I did
 R: What about extraversion?
 O: on extroversion, you scored pretty close to the median
 R: Do you know the link with music features?
 O: in the explanations chart there is a *strong relationship
 between extroversion and dancability and liveliness*
 O: but if I'm reading this correctly, your preferences didn't
 indicate strongly one way or another about those features
 (D 13)

In excerpt 4.15, O first highlights a warrant (a general relationship between openness and acousticness), and then links the warrant with the assessment (that R is very open) through a datum (R's preference for acoustic music). A similar strategy is observed in the excerpt below, where O's generic statement concerning the preferences that yield a high agreeableness score can be characterised as a warrant (or a conjunction of warrants):

- (4.16) R: Why do you think I'm a very agreeable person?
 R: * disagreeable
 O: I don't know, but the results says that *a person scores high in
 that category if they prefer non-live music, major mode, acoustic
 music, and non-instrumental music ...*
 (D 8)

A related behaviour can be observed in excerpt 4.17 below, where O explains an assessment concerning a specific trait by elaborating different sub-traits:

- (4.17) O: You have been rated the highest in openness
 R: Oh wow, why? :)
 O: Openness describes a dimension of style that is imaginative
 O: creative people
 O: down to earth
 O: and conventional
 R: are these three different grades?
 O: Agreeableness is rated 0.5
 O: The reasons that I gave was all for openness (D 15)

Although the conveyed sub-traits (imagination, creativity, etc.) are presented to operators in the tool as information concerning what personality traits *mean*, O in excerpt 4.17 treats—and explicitly refers to—these sub-traits as *reasons*. In a possible interpretation, O implicates that the AI generally associates openness with imagination, creativity, etc., and that the AI assesses R to be imaginative, creative, etc. Note, however, that in this case the implication is false, since the AI does not assess traits in terms of sub-traits.

With respect to modelling, all observed types of explanantia are deemed potentially relevant. However, the algorithmic strategy (excerpt 4.9) is excluded primarily due to being informationally excessive, which can be assumed to increase cognitive load for users. It is also linguistically less interesting to model, given its relatively straightforward relation to numerical model information. Consequently, algorithmic narratives are rewritten to another, e.g. argumentative, strategy, as in the example below:

- (4.9') U: Why do you think I score high on openness?
 S: You seem to prefer silent music.

While the ability to provide the source of evidence or to mention factors without detailing their relation to the particular judgement are potentially relevant for CXAI, it is here assumed that, as data is concerned, explainees are typically interested in more informative explanations, rather than informationally vague explanations (unless framed in ways that implicate relevant information; see section 4.5.5). This is supported by the observation that in cases where O provides general explanations, specificity is increased later in the same dialogue. Furthermore, positive feedback from R (such as “mmm ic” and “ok, I see”) is only observed following more informative explanations. Therefore, informationally vague datum explanations are rewritten to more informative ones.

4.5.4 Response strategies

In some cases, O's choice of response strategy aligns directly with R's mode of query. This is particularly evident in the case of polar/confirmation questions, as in the excerpt below:

- (4.18) R: But the result could be because of a preference of music
with lower tempo
R: am I correct?
...
O: Yes this is correct. You preferred music with lower tempo.
(D 2)

In excerpt 4.18, the polar nature of R's query is explicitly reflected in O's affirmation ("Yes this is correct"), while the potential explanans embedded in R's question (a preference for music with low tempo) is explicitly confirmed by echoing it back.⁴ However, in many other cases, the degree of alignment between query type and response strategy is weak or unclear. For example, in excerpt 4.8, where O mentions factors, R had previously asked: "Why is it a 5 on Openness, based on *what variables*". On the other hand, O's initial response to R's query used an algorithmic narrative (see excerpt 4.9), possibly triggered by the quantitative nature of O's highlighted explanandum ("Why is a 5 on *Openness*"), although such a link is far from evident. Furthermore, many queries in the corpus are constructed using open-ended wh-questions such as "*what* do you base this conclusion on" or "*Why* do you think I'm a very agreeable person?" and thereby do not clearly indicate an expected explanans mode.

With respect to dialogue modelling, the lack of evident patterns regarding choice of response strategy for open-ended wh-questions provides limited guidance. The path taken here is to assume that in situations where the user's query does not indicate a particular kind of desired explanans, the system has some means of selecting a response strategy (e.g. a default strategy).

⁴While O's confirmation in excerpt 4.18 paraphrases the potential explanans embedded in R's question, it does so by merely echoing the explanans as such, without conveying it as an explanans in relation to the explanandum at hand. Such a strategy may be ambiguous or even misleading in the sense that a potential explanans might be true (yes, R indeed prefers low-tempo music) without being related to the explanandum (no, R's preference for low-tempo music is not a basis for the AI's assessment). For this reason, O's response is written to "Yes, your preference for low-tempo music supports the assessment that you score low on extraversion."

4.5.5 Argumentative structure

As discussed above, the epistemic basis of the AI’s judgements are frequently explained either via data (e.g. a preference for acoustic music), warrants (e.g. a relationship between acousticness and openness), or a combination thereof. Argumentatively, conveying only data or warrants to support a claim (in this case concerning a person’s personality) can be understood as an *enthymeme* (Breitholtz 2020), i.e., a truncated argument. Through this theoretical lens, a conclusion can be comprehensible if one or more premises that would make the argument more complete has been omitted, to the extent that the receiver is able to identify the missing premise(s). This can potentially explain why respondents sometimes signal understanding of operators’ seemingly truncated explanations, as in the example below:

- (4.19) O: you have scored low on “openness” and “neuroticism”
O: slightly higher on “extraversion”
...
R: I really want to know what these results are based on ...
O: “openness to experience”: the score was calculated on the basis that you don’t prefer high-energy, loud music
R: well, I would not consider any of the music pieces I listened to as high-energy and loud music...but *ok, I see*
...
O: “Conscientiousness” was based on the fact that you prefer non-live music and music with spoken words
R: *mmm ic*
O: “Extraversion” is based on the fact that you prefer music with spoken words while you disprefer sad/depressed/angry music, non-instrumental music, non-danceable music
R: *yes, this sounds about right*
O: “Agreeableness” is based on the fact that you prefer non-instrumental music, non-live music, and perhaps acoustic music
...
R: agreeableness’s *explanation also sounds good to me*
(D 19)

In excerpt 4.19, O’s explanations are consistently enthymematic, in that they explain claims merely via data. Potential warrants that would make the explanations more complete, e.g. that the AI generally associates a dispreference for high-energy music with low degrees of openness, are not verbalised—even if

such information is available to operators in the tool. Nevertheless, R consistently signals acceptance / positive understanding in response to O's explanations, indicating that R is able to infer implicit premises.⁵

One factor that might explain why respondents seem to understand operators' enthymematic explanations is that data are conveyed in ways that indicate which kind of warrant might be at play. Specifically, operators use numerically vague datum expressions such as "you don't prefer high-energy music", despite having access to more detailed visualisations of the respondent's music preferences in the tool (see figure 4.4). Presumably, the simple polar form of these expressions implicates a monotonic relation between the feature at hand and the target, e.g. that a preference for low-energy music is associated with low openness. If, instead, O had explained R's low estimated openness by describing R's preference for energy as "around 30 on a scale from low energy (0) to high energy (100)", this might have raised the question of whether a value of 30 specifically is indicative of low openness. This finding resonates with similar observations in the non-AI-assisted medical dialogues in chapter 3 (section 3.4.2).

In terms of modelling, the approach taken here is to let the system explain predictions with either data or warrants, depending on some policy. (The policy as such is not modelled.) Consequently, datum-based strategies are rewritten as follows:⁶

- (4.10') U: I really want to know what the result is based on...why
am I so low on openness? kind of disagree with that
S: You seem to prefer low-energy, loud music.
...

Warrant-based strategies are rewritten as in the example below:

- (4.15') S: I think you rate high on openness.
U: It's interesting, I wonder why
S: Openness is associated with a preference for acoustic
music.
...

⁵In an alternative interpretation, R's positive feedback merely affirms the validity of the data as such, rather than their explanatory function in relation to the AI's judgement. For example, when R utters "yes, this sounds about right", this might primarily affirm that R indeed prefers music with spoken words, etc. However, feedback such as "ic" (I see) and "explanation also sounds good to me" seems to affirm the explanation as a whole. Possibly, if R had deemed the data valid but unintuitive in relation to the AI's judgement, R would not affirm O's explanations in this manner.

⁶Methods for extracting datum- and warrant-based explanations for predictions from predictive models are discussed in chapter 5.

4.5.6 Faithfulness and sources of evidence

In most cases where operators explain the AI's predictions, pieces of evidence accurately reflect the information in the tool and are thereby grounded in the actual workings of the model. However, in several instances, operators' reasoning goes beyond the information in the tool, as in excerpt 4.12 (repeated below for convenience):

- O: You have high on openness, and very low on neuroticism. neutral on the rest
R: oh, low on neuroticism, but I pressed so fast on the growling-dislike-button!
...
O: *Maybe if you were neurotic you would not like the growling, or you could. I can see reasons in both directions*

In excerpt 4.12, O's reflections regarding the potential relation between neuroticism and preference for growling are quite evidently O's own, and not grounded in the information in the tool. This can be concluded from the fact that the tool does not convey how ratings of specific tracks (such as the one with growling) influence predictions. Furthermore, operators can deduce from the information in the tool that growling is not included among the features (music preferences) considered by the model.

Similarly, in excerpt 4.11 (repeated for convenience below), O implicates a relation between neuroticism and preference for death metal which is not grounded in the model:

- O: Your neuroticism was -1.7
...
O: Do you not enjoy death metal
R: not particularly
R: but slightly
O: Fair enough
O: Explains the slightly negative score

In terms of dialogue modelling, it is here assumed appropriate for the explainer to only provide information that a CXAI system would realistically have access to. Consequently, when a user explicitly refers to a concept that goes beyond the system's information about the model or the domain, as in excerpt 4.12, the system signals its inability to answer the question (see section 4.5.7 below). Occurrences where an operator unpromptedly explains a prediction in a model-ungrounded way

(excerpt 4.11) are not rewritten, under the normative assumption that it is better to not explain a prediction unprompted than to provide a model-ungrounded unprompted explanation.

4.5.7 Answer unavailability

When operators cannot provide an answer to an explanatory question raised by the respondent, various different strategies are used. One such strategy is to *answer a related question*, which can be understood as adhering to the maxim of relevance (Grice 1975). In the following excerpt, instead of answering R's question about the contribution of specific ratings, O answers a question concerning the meaning of one of the traits at hand:

- (4.20) R: ... Among my answers, which ones tilted me the most towards being non-agreeable and slightly open?
O: *Your agreeableness reflects individual differences in concern with cooperation and social harmony.* (D 3)

Another strategy is to *signal answer unavailability*, which can be understood as adhering to the maxim of quality (Grice 1975), as in the following excerpt:⁷

- (4.21) R: So, what did they base the low conscientiousness on?
O: *hard to say, any specific music you rated high or low* (D 33)

Similarly, in excerpt 4.5 (repeated below for convenience), O aligns with R's interrogative stance (towards the causal relationship between music listening and personality), thereby implicating a lack of knowledge concerning the question at hand:

- R: I wonder if music influences the personality or if it's only the other way
O: *yeah I was thinking that too*

In some cases, answer unavailability signalling is combined with *offering an answer to a related question*, as in excerpt 4.15, repeated below for convenience:

- R: It's interesting, I wonder what song would give this trait
O: well I actually can tell you something about that I think
O: *not which song in particular, but how openness relates to features of the music*

⁷O's phrase "any specific music you rated high or low" might be understood as a vague/partial answer to R's question since it alludes to the kind of information considered by the AI. Nevertheless, O's overall rhetoric seems to convey that no satisfactory answer can be obtained.

R: Oh great I'm interested

A similar combination of strategies can be observed in the following excerpt:

(4.22) R: So there were no specific answers in the test that led to these results? For instance my slight disliking of death metal.

O: *Not really, your answer for the tracks is not something i can see, What i can see is What your answers acumilated to. Then showing me What personality traits you have*

R: I see, that makes is a lot clearer. (D 3)

In excerpt 4.22, R's response to O's answer unavailability feedback strongly conveys positive understanding, indicating that answer unavailability feedback can help explainees form a correct mental model of how the AI makes its predictions.

Among the observed strategies, signalling answer unavailability is here considered the most important for dialogue modelling, since it helps users form correct mental models. Assuming that the system is able to parse (at least some) questions for which answers are unavailable, exchanges can be rewritten as follows:⁸

(4.20') U: Among my answers, which ones tilted me the most towards being non-agreeable?

S: *I don't have any information about that.*

(4.5') U: I wonder if a preference for acoustic music influences openness or if it's only the other way

S: *I don't have any information about that.*

The rewritten excerpts 4.20' and 4.5' assume that the NLU component can identify and represent relations concerning which the system lacks information. However, it is unrealistic to assume that *any* unanswerable question that a user might ask can be represented by the NLU using the system's internal semantics. For example, music concepts such as "death metal" and "growling", whose semantics are not related in any evident way to the features or target labels at hand, are here assumed unparsable by the NLU. Consequently, when excerpts where respondents ask questions involving such concepts are rewritten, the system gives negative understanding feedback:

⁸In principle, the influence of specific ratings on predictions (rewritten excerpt 4.20') *can* be explained for the model at hand, given the interpretable relation between specific ratings and features (see section 4.2). However, such information was not provided to operators in the interface.

- (4.12') S: I think you rate low on neuroticism.
 U: oh, is this because I don't like growling?
 S: *Sorry, I don't understand what "growling" means in this context.*
- (4.11') U: do you think death metal fans are 5 or -5 in neuroticism?
 S: *Sorry, I don't understand what "death metal" means in this context.*
- (4.22') U: So there were no specific answers in the test that led to these results? For instance my slight disliking of death metal.
 S: *Sorry, I don't understand what "death metal" means in this context.*

4.5.8 Feedback and grounding

Interlocutors (especially respondents) use various forms of feedback such as "ok" and "I see", as in the following excerpt:

- (4.23) O: the AI calculates the results based on a statistical model for each personality trait
 O: I doubt if it can discriminate different phases
 R: *hm ok*, so it takes into account many other people's statistics then
 O: of a particular person
 R: yes, what you described is a very general picture
 O: a 1000 users apparently
 R: *aha I see*, quite a small set
 R: ok, I am *pretty satisfied with the explanation*
 O: users of the music website Last.fm
 O: I am glad :) (D 19)

In excerpt 4.23, O either accommodates R's acknowledgement silently, or positively affirms it. Another observed behaviour in response to acknowledgements from R is to continue to the next sub-topic or next piece of information to deliver, as in excerpt 4.19 (repeated for convenience below):

- O: you have scored low on "openness" and "neuroticism"
 O: slightly higher on "extraversion"
 ...
 R: I really want to know what these results are based on ...

- O: “openness to experience”: the score was calculated on the basis that you don’t prefer high-energy, loud music
- R: well, I would not consider any of the music pieces I listened to as high-energy and loud music...but *ok, I see*
- ...
- O: “Conscientiousness” was based on the fact that you prefer non-live music and music with spoken words
- R: *mmm ic*
- O: “Extraversion” is based on the fact that you prefer music with spoken words while you disprefer sad/depressed/angry music, non-instrument music, non-danceable music
- R: *yes, this sounds about right*
- O: “Agreeableness” is based on the fact that you prefer non-instrumental music, non-live music, and perhaps acoustic music
- ...
- R: agreeableness’s *explanation also sounds good to me*

Negative understanding feedback from R can also be observed, as in the following excerpt (continuation of excerpt 4.19):

- (4.24) R: I think these results kind of make sense for myself, but *I am having hard times understanding my openness rating.* It’s interesting that I am extravert but not open based on these results...don’t you see a confusion there?
- R: agreeableness’s explanation also sounds good to me
- R: i see now
- O: “Neuroticism” is based on the fact that you disprefer acoustic, high-energy music while you slightly prefer non-danceable music
- ...
- (D 19)

In excerpt 4.24, negative understanding feedback yields no evident response from O, possibly due to the presence of multiple parallel issues being raised. Another observed response to negative understanding feedback from R is to treat the feedback as a matter of agreement:

- (4.25) O: Do you agree or disagree with the results
- R: but *it is hard to understand* what in the music made the AI think that I am 0 in extraversion...
- O: ?
- O: Yeah, so disagree?
- (D 15)

In excerpt 4.25, a possible reason for O's behaviour is that O does not know how to explain the AI's judgement. This hypothesis is supported by the fact that, throughout the interaction, this operator does not provide any specific explanations for the AI's judgement. With respect to dialogue modelling, none of the observed response behaviours above seem evidently ideal to emulate from a normative perspective. Instead, it may be more adequate to treat negative understanding feedback as requests for an explanation, and hence, to provide an explanation (if such is available).

Feedback concerning presupposition violations is also observed, as in the following excerpt, where O signals that the presupposed content embedded in R's why-question is false:

- (4.26) R: Canyou give me my results
O: opennessa 5
...
O: neurotisk -1,2-ish
...
R: Why am i neurotisk
O: *no, minus 1,2*
R: Which means i am not neurotisk
O: guess so (D 10)

In another case, R asks a question which seems to presuppose that R has other music preferences than those inferred by the AI from R's ratings, which leads O to highlight the mismatch:

- (4.27) R: ... I listen to a lot of heavy metal. Is there a correlation
between loud and fast music and my score?
O: *You seem to prefer slower and less loud music*
R: haha that's the complete opposite. ... (D 8)

With respect to modelling, it is assumed the NLU component is able to identify presuppositions in user input, and that the dialogue manager signals potential presupposition violations concerning, e.g., predictions or feature values embedded in questions.

4.5.9 Anaphora

Anaphoric references to explananda, such as "what do you base *this conclusion* on", are frequent in the analysed empirical material. As in chapter 3, anaphoric references are assumed to be contextually resolved by the NLU component.

4.5.10 Turn-taking and complex explanantia

Interlocutors frequently combine multiple pieces of explanantia in a single utterance or sentence. For example, in excerpt 4.10, O explains the AI's assessment concerning extraversion by combining multiple music preferences using a contrastive structure in a coordinating conjunction: "Extraversion' is based on the fact that *you prefer music with spoken words while you disprefer sad/depressed/angry music, non-instrument music, non-danceable music*". Elsewhere in the same excerpt, coordination takes place on the adjectival level, when the operator states that "the score was calculated on the basis that *you don't prefer high-energy, loud music*".

In contrast, in excerpt 4.17 (repeated in part for convenience below), O delivers sub-explanations incrementally across multiple utterances:

- O: You have been rated the highest in openness
- R: Oh wow, why? :)
- O: *Openness describes a dimension of style that is imaginative*
- O: *creative people*
- O: *down to earth*
- O: *and conventional*

It is also worth noting that O uses the discourse marker "and" to explicitly mark the final utterance as a topic continuation.

In some cases, acknowledging feedback from the receiver seems to serve as a signal to provide additional information, as in excerpt 4.19 (repeated in part below for convenience):

- O: "Conscientiousness" was based on the fact that you prefer non-live music and music with spoken words
- R: *mmm ic*
- O: "Extraversion" is based on the fact that you prefer music with spoken words while you disprefer sad/depressed/angry music, non-instrument music, non-danceable music
- R: *yes, this sounds about right*
- O: "Agreeableness" is based on the fact that you prefer non-instrumental music, non-live music, and perhaps acoustic music
- ...

With respect to modelling, it is here assumed that an interlocutor who delivers complex information has some means of designing a suitable turn-constructive unit (Sacks, Schegloff and Jefferson 1974) at any given moment in the dialogue. For example, an interlocutor may have knowledge about two explanantia, but

decide to initially only mention one of them, as exemplified by the rewritten excerpt below:

- (4.10'') U: I really want to know what the result is based on...why am I so low on openness? kind of disagree with that
S: You seem to prefer low-energy music.
U: ok
S: Also, you seem to prefer loud music.
...

4.5.11 Ellipsis and explanandum co-referencing

When operators respond to explanation requests, they occasionally omit the explanandum targeted by the respondent, presumably since it is evident from the dialogue context, as in excerpt 4.26 (repeated in part for convenience below):

- R: Why am i neurotisk
O: *no, minus 1,2*
R: Which means i am not neurotisk
O: guess so

In excerpt 4.26, O's highlighted response is *elliptical* and (correctly) assumes that R can infer from context that the target of the feedback is R's question about neuroticism.

In other cases, in their responses, operators re-state the explanandum targeted by the respondent, as in excerpt 4.10 (repeated in part for convenience below):

- R: I really want to know what these results are based on...why am I so low on openness? kind of disagree with that
O: "*openness to experience*": the score was calculated on the basis that you don't prefer high-energy, loud music
...

In some cases, co-references to previously raised explananda are anaphoric rather than explicit, constituting a middle ground between ellipsis and explicit re-statement, as in the excerpt below:

- (4.28) O: ... You scored 5 in Openness, a bit under -2 in Conscientiousness, a bit over 1 in Extraversion, a bit under -4 in Agreeableness, and almost -3 in Neuroticism
- R: Why do you think I'm a very agreeable person?
- R: * disagreeable
- O: I don't know, but the results says that a person scores high in *that category* if they prefer non-live music, major mode, acoustic music, and non-instrumental music. Maybe you have other preferences (D 8)

Presumably, the choice concerning whether to respond elliptically depends on the extent to which the operator assumes that the respondent can infer the explanandum from context. However, the conditions influencing this choice cannot straightforwardly be determined merely by analysing the empirical material at hand.

The data suggests that choice of co-referring strategies is not only influenced by considerations concerning contextual salience, as discussed above, but also serves to interactively coordinate how predictions might be interpreted. In excerpt 4.26, interlocutors refer to the prediction at hand both numerically (“minus 1,2”) and polarly (“Which means i am not neurotisk”). A similar pattern is observed in excerpt 4.28 (“a bit under -4 in Agreeableness” / “very agreeable person”). As exemplified by excerpt 4.26, an interpretation by one interlocutor can be denied/corrected by the other interlocutor. With respect to modelling, the path taken here is to let the system consistently deliver its assertions polarly (with potential hedging), which reduces the need to coordinate numerical interpretations. However, in future work, it might be relevant to further explore semantic coordination in relation to co-referring strategies in the context of XAI (see section 6.5.4).

4.5.12 Sub-topic structure

Issues concerning predictions and their bases are frequently raised on a general level, without highlighting a specific trait. The general topic is then broken down into several sub-topics corresponding to specific traits. For example, in excerpt 4.29, the respondent first raises a general question regarding “test results”, which the operator then addresses on sub-topic level:

- (4.29) R: May I have my test results?
 ...
 O: Judging from What i can see you are not very agreeable
 O: However you are slightly open
 R: Bold statements! What do you base them on ... (D 3)

Since management of sub-topic structure as a dialogue phenomenon is here deemed unrelated to explanatory communication strategies, sub-topic structure is flattened during rewriting. For example, excerpt 4.29 is rewritten as follows:

- (4.29') U: What do you think regarding my agreeableness?
 S: I think you rate quite low on agreeableness. ...

4.5.13 Reliability and epistemic stance

Interlocutors use different strategies for managing reliability of the AI's judgements and coordinating epistemic stance in relation to them. One such strategy focuses on highlighting overall limitations of the AI's reasoning capabilities, as exemplified by the following excerpt:

- (4.30) O: the AI calculates the results *based on a statistical model* for each personality trait
 O: *I doubt if it can discriminate different phases*
 R: hm ok, so it takes into account *many other people's statistics* then
 O: *of a particular person*
 R: yes, what you described is a *very general picture*
 O: a *1000 users* apparently
 R: aha I see, quite a *small set*
 ...
 O: *users of the music website Last.fm*
 ...
 R: *very specific website—most of the people right now don't really use Last.fm, maybe that would explain a lot*
 ...
 O: I agree (D 19)

A similar strategy can be observed in the excerpt below:

- (4.31) O: I can see that you have very high scores on openness
 R: That is to be expected
 R: *Selection bias for people that participate in this kind of experiment*
 R: *or any experiment* (D 33)

In Toulmin’s argumentation framework (Toulmin 2003), this strategy focuses on how warrants are *backed*. In terms of dialogue modelling, similar behaviour can potentially be emulated by providing the system with information concerning, e.g., model type and nature of training data.

Another strategy involves *epistemic detachment* from the AI, achieved with hedging markers such as *perhaps* (“Agreeableness’ is based on the fact that you prefer non-instrumental music, non-live music, and *perhaps* acoustic music”), stance markers such as “*apparently* acousticism is positively correlated with neuroticism” and metadiscursive expressions such as “*if I’m reading this correctly*, your preferences didn’t indicate strongly one way or another about those features”. In excerpt 4.28 (repeated for convenience below), O’s epistemic detachment takes the form of explicit evidential marking (“*I don’t know, but the results says ...*”) and occurs in response to R’s presupposed attribution of the AI’s judgement to O:

- O: ... You scored 5 in Openness, a bit under -2 in Conscientiousness, a bit over 1 in Extraversion, a bit under -4 in Agreeableness, and almost -3 in Neuroticism
 R: Why do *you* think I’m a very agreeable person?
 R: * disagreeable
 O: *I don’t know, but the results says* that a person scores high in that category if they prefer non-live music, major mode, acoustic music, and non-instrumental music. Maybe you have other preferences

With respect to dialogue modelling, the path taken here is to hedge the system’s predictions based on confidence estimates (see section 6.3.1). For example, a moderately confident prediction can be expressed as “I think you rate low on openness” or “It seems that you have an open personality”, without explicitly representing propositional attitudes—in this case, that the belief concerning R’s personality is ascribed to the system itself—in the internal semantics of the system. In contrast, modelling the type of interactively coordinated management of epistemic stance observed in excerpt 4.28 would require propositional attitudes to be represented (see section 6.5.6).

4.6 Limitations

Since the models used in the experiment are employed differently in inference than in training, their predictive performance cannot straightforwardly be evaluated using conventional metrics such as accuracy. However, this limitation was deemed to have no substantive impact in relation to the focus of the study.

While English proficiency levels are generally high in the local population, potential language deficiencies might still have affected the dynamics and content of the collected dialogues, although the analysis did not reveal any evident instances to that effect.

In trials 3–5, data collection did not achieve its theoretical potential in terms of amount of collected data. This is reflected in the frequencies for these trials in table 4.1, where there is a discrepancy between, on the one hand, the number of dialogues that theoretically could have been collected given the number of participants and the number of interactions per participant and, on the other hand, the actual number of collected dialogues. For example, in trial 3, the total number of participants (15) and interactions per participant (2) could in principle enable $7 \cdot 2 = 14$ dialogues to be collected (since 7 of the 15 participants could be paired up with another participant twice); however, only 12 dialogues were collected. The reasons for this discrepancy were that three participants did not proceed through all the steps in the experiment, and that one pair of participants did not exchange any chat messages. The same set of failure modes were observed for trials 4 and 5.

In some instances, it was observed that participants were not paired up as expected due to technical problems. While this explains why some participants did not complete the entire experiment, the exact reasons for each incomplete participation could not be fully determined; possibly, some participants simply aborted.

It can also be observed that a relatively low ratio (26%) of collected dialogues was revealed to address topics of interest to the present study. Although the low ratio is not necessarily problematic by itself, combined with a relatively limited sample it resulted in a small amount of data of interest. Issues related to generalisability will be further discussed in section 7.2.

4.7 Discussion and conclusions

The dialogue behaviours observed in the present chapter share many similarities with the non-AI-assisted medical dialogues analysed in chapter 3 when it comes to aspects such as explanation queries, types of explanantia, turn-taking, and

anaphora. By itself, this finding suggests that humans might use and prefer the same kind of dialogue strategies, regardless of whether the judgements whose epistemic bases are addressed emanate from a predictive model or the interlocutors own minds. For example, while one of the operators used an algorithmic narrative to convey how a certain prediction was calculated, in many other instances, the epistemic basis was addressed argumentatively and using enthymemes, in ways that resemble the strategies observed in chapter 3.

One important difference in relation to chapter 3 is that the access to ground truth epistemic bases for the AI's predictions makes it possible to assess how the observed human explanations relate to the actual reasoning used by the AI. In this regard, it was observed that while evidence stated by explainers tend to accurately reflect information presented to them in the tool, and thereby also the actual workings of the AI, in some cases explainers go beyond the model-grounded information. In relation to CXAI, this observation highlights an important design choice concerning whether human–AI interfaces should focus entirely on explanations grounded in the actual workings of the AI, or potentially also offer explanations obtained from other knowledge sources. Given this thesis' normative focus on transparency, the path taken here has been to focus on model-grounded explanations when rewriting human–human to human–AI dialogues. This can be contrasted with the systems proposed by Feustel et al. (2024) and Schindler et al. (2025), who augment model-based explanations with domain knowledge derived by other means. However, as will be discussed in section 5.1.4, since such explanations are not grounded in the workings of the predictive model, they can be potentially misleading with respect to why a particular model makes a certain prediction. It is for this reason that the present study focuses on model-grounded explanations. An alternative approach for augmenting model-grounded explanations with external domain knowledge, while maintaining transparency, could be to let the system use both kinds of knowledge and state the source of evidence. For example, if a user asks why a short loan duration increases the chance of having one's credit application approved, the system can clarify that the model can only observe statistical correlations in data, and not the potential causes behind these correlations. It can also offer a causal explanation for this correlation, while making clear that this explanation might not reflect the actual reason for why the correlation exists in the training data.

The analysis also highlights how framing of explananda shapes explanatory interaction. Specifically, when explainers convey the AI's predictions in their own words to explainees, they need to interpret and frame the graphically presented information, e.g. in numerical or more evaluative terms. The study highlights how this process contributes to shaping the construction of the explanandum and thereby plays a central role in explanatory interaction. For example, when

the operator in excerpt 4.3 conveys that, apparently, the respondent is “very open”, and the respondent wants to know “what song would give this trait”, then, in some sense, the explanandum targeted by the respondent has already been constructed by the operator. Importantly, this construction involves an element of subjectivity which goes beyond the raw quantitative information conveyed by the statistical model. Methodologically, this distinguishes the present study from the one in chapter 3, where the analyst cannot observe the mental representations on which the explainers’ verbal judgements are based. For example, when the doctor in excerpt 3.3 expresses that a lung transplant isn’t feasible, the analyst lacks information concerning whether this assessment is based, e.g., on a specific estimated probability of successful outcome which the doctor deems as too low. (The doctor may not himself have conscious access to such mental representations.) The framing of explananda raises issues concerning design of CXAI systems. If the system expresses its predictions in evaluative terms, as frequently observed in the collected human dialogues, some means of classifying raw outputs from the predictive model is required, e.g. by dichotomising outputs in relation to a median instance or a pre-determined threshold. However, such design choices can undesirably shape users’ understanding by anchoring their perception of the prediction to the system’s interpretation (Tversky and Kahneman 1974; Scurich 2018; Krauss, G. I. Cook and Klapatch 2018). It should also be noted that even if the system conveys predictions numerically (e.g. “3 on openness (on a scale from -5 to 5)”), users may still frame explananda evaluatively (e.g. “why do I score so high on openness?”), thereby presupposing an interpretation which the system needs to either accept or challenge. This is particularly pertinent in situations where it might be relevant for the system to flag presupposition violations, as in excerpt 4.26, where the respondent seems to misinterpret the numerically conveyed information. This suggests that the system may need to be able to interpret raw model outputs, even if communication of such interpretations can have potential negative consequences.

In relation to issues concerning faithfulness and sources of evidence, the chapter revealed various ways through which interlocutors signal their epistemic stance in relation to the predictions and evidence. These observations indicate design choices concerning how CXAI systems might position themselves epistemically in relation to the predictive model at hand. The approach taken here has been to manage epistemic stance on the level of surface realisation, rather than semantically. In practical terms, epistemic positioning is thereby delegated to the NLG component of a dialogue system. An alternative approach could be to represent epistemic stance on the level of dialogue management. This would, in principle, enable epistemic stance to be addressed and coordinated interactively in human–computer dialogue. Such a capability can be important from

the perspective of accountability and trust in critical domains such as healthcare, where transparency concerning the source of a prediction or explanation can be important. For example, if a patient asks a conversational healthcare system why “it” predicts a certain outcome of surgery, it might be important that the system clarifies the actual source of the prediction (namely a predictive model), even if the source was not explicitly interrogated by the user (as in excerpt 4.28). Such a behaviour hinges on an ability to detect epistemic presupposed agency in user input and to highlight the actual source of a claim or evidence in the case of a presupposition violation.

4.8 Ethics declaration

Participation in the study was anonymous and no personal data such as participant’s names were collected. Dialogue data was screened before being stored to ensure that collected utterances did not contain personal data. To reduce risk of over-reliance on AI predictions, participants were informed in a debriefing that the method used in the experiment for estimating personality from music preferences is experimental and has not been validated scientifically. Consent from participants was obtained in the form of a checkbox on the experiment’s start page, through which participants agreed that their interaction would be collected, stored and made available for research purposes.

The Swedish Ethical Review Authority issued an advisory opinion (case number 2022-06881-01) stating that since the study does not involve any physical or other intervention on participants, as covered by Section 4 of the Swedish Ethical Review Act, or any processing of personal data of the kind covered by Section 3 of the Ethical Review Act, the study is not covered by the provisions of Sections 3–4 of the Ethical Review Act and therefore should not undergo ethical approval. The advisory opinion also stated that the authority raises no ethical objections to the research project.

CHAPTER 5

Extracting arguments from interpretable models

One of the main findings in chapters 3 and 4 was that humans often explain judgments argumentatively, i.e., as evidence supporting a claim, and enthymematically, i.e., by presupposing certain argumentative elements rather than communicating them explicitly. For example, a claim may be supported by case-specific evidence (data) without explicitly conveying general rules or principles (warrants) that link the evidence to the claim.

The present chapter investigates how the argumentative and enthymematic nature of observed human explanations can be emulated in linguistic AI explanations for model predictions. While previous work in XAI has primarily focused on how various kinds of explanatory information can be extracted from ML models, argumentative and linguistic aspects of AI explanations have received much less attention. Specifically, prior research has not investigated how AI explanations can be generated in ways that align with typical analogous human explanations, while also inviting accurate inferences concerning the nature of the model's reasoning. The present chapter addresses this gap in previous work by proposing a method for extracting arguments from a particular class of interpretable models, namely generalised linear models (GLMs), providing formal guarantees for correctness of invited inferences, and applying the method in a case study involving AI-assisted decision-making.

The remainder of the chapter is organised as follows. Section 5.1 introduces important concepts and positions the work in relation to previous research. A novel method for extracting arguments from GLMs is proposed in section 5.2. In section 5.3, it is shown that enthymematic explanations generated using the proposed method naturally invite inferences that faithfully reflect that actual reasoning steps of the model. The proposed method is then applied in a case study (section 5.4) involving predictions of outcome of surgery in the context of spinal surgery decision-support. Findings from evaluating the proposed approach through a design workshop with orthopaedic surgeons are also presented. Finally, section 5.6 discusses findings and offers some conclusions.

5.1 Background and related work

This section discusses and positions the work in the present chapter in relation to previous work on enthymemes, users' interpretation of AI explanations and argument extraction from ML models.

5.1.1 Enthymemes

The argumentative incompleteness of enthymemes raises questions concerning how enthymemes are understood by hearers. Breitholtz (2020) proposes that comprehension of enthymemes hinges on the addressees' identification of presupposed argumentative elements under the influence of prior knowledge. Breitholtz (2020, p. 18) discusses the following example:

- (5.1) A: Let's walk along Walnut Street. It's shorter.
(Walker 1996)

According to Breitholtz, the addressee in excerpt 5.1 makes sense of how the argument (that Walnut street is shorter) is presented in support of the claim (that the route along Walnut Street is favourable) by *assuming that A presupposes that short routes are generally favourable*. Generally, for this kind of presupposition to be perceived as coherent and relevant by the addressee, it needs to match the addressee's prior beliefs and knowledge (Breitholtz 2020, p. 21). Addressees do not arbitrarily comprehend and accept *any* enthymematic explanation; the "hole" in the argument needs to be filled with a piece that not only fits the hole; it also needs to be easily identified in the addressee's cognitive resources and combine well with other pieces of beliefs and knowledge in those resources.

Following Ducrot (1988), Breitholtz uses the Aristotelian term *topos* (pl. *topoi*, roughly corresponding to Toulmin (2003)'s term *warrant*) to refer to kind of general knowledge that "fills the gap" left by enthymemes. According to Ducrot (1988), *topoi* tend to be *gradual*, by which is meant that they presuppose a monotonic relationship between two scalar predicates (e.g., a walking path's length and the extent to which it is preferred). Gradual *topoi* are more applicable for extreme predicate values; for example, a *topos* concerning preference for short paths is more applicable the shorter the path is.

Ducrot's theory resonates well with the frequent use of scalar triggers observed in human-human explanatory dialogues in chapter 3, where expressions such as *lower* (levels of oxygen in the blood than one would expect) and *over* (two months of medication use) seem to suggest a monotonic relation between the predicates at hand.

While Ducrot argues that presupposed relations in discourse between predicates tend to be monotonic, he acknowledges that, in some cases, both a topos and its “inversion” can potentially be used simultaneously by a speaker in a single utterance (Ducrot 1988, p. 4). For example, if speaker B wants to get some exercise, she may respond that the path along Chestnut Street is “too short”, although B generally shares A’s preference for shorter routes.

The present chapter primarily focuses on Ducrot’s main setting where two predicates are assumed to be associated with a single gradual (monotonic) topos, and uses this setting in its analysis of how AI users might interpret enthymematic AI explanations, as well as in its proposed method for generating such explanations. More elaborate scenarios involving non-monotonic relationships and/or simultaneous use of mutually “inverted” topoi will briefly be discussed in section 5.6.

5.1.2 Explanation faithfulness

As discussed in chapter 2, AI explanations should not only be comprehensible for users, but also accurate (faithful) with respect to how the AI actually makes its predictions. In relation to enthymemes, this constitutes a potential source of concern. In principle, an AI system may convey true and relevant evidence in support of its predictions, while simultaneously misleading users to form incorrect beliefs about the AI’s inner logic. To illustrate this problem with a concrete scenario, we can conceive that a hypothetical bank secretly uses the following, somewhat peculiar, algorithm to make decisions concerning loan applications:

1. Calculate the sum of all the digits in the applicant’s annual income (in euros).
2. If the digit sum is even, grant the application; otherwise reject the application.

For example, if Sam earns €31,000 per year, then the digit sum is $3+1+0+0+0 = 4$, which is an even number, so the loan is granted. To explain decisions, the bank uses the following statement: “Your application has been granted/rejected on the basis of your income.” This statement is always true and relevant for any decision, in the sense that the algorithm *does* base its decisions on the applicant’s income. However, applicants might make sense of decision statements by drawing pragmatic inferences which lead them to forming incorrect mental models of how the bank generally makes its decisions. For example, customers with relatively high income whose applications are granted might interpret the bank’s statement as indicating that their income is deemed sufficiently high. Conversely, customers

with relatively low income whose applications are rejected might interpret the bank’s statement as indicating that their income is too low. In fact, however, both inferences are incorrect, since the algorithm does not assess income in low/high terms.

Importantly, as the present chapter will show, the problem with potentially misleading AI explanations does not only concern hypothetical contrived algorithms, but also conventional ML models explained with popular XAI approaches. Common methods for explaining predictions by opaque models cause similar problems as in the thought experiment, and for similar reasons, namely a combination of vague explanations and an unintuitive or inscrutable AI reasoning. One of the main motivations behind the proposed method is to mitigate the risk of what might be called an *illusion of explanatory grounding*, i.e., a misperception that the epistemic basis for a judgement has been grounded (mutually established between interlocutors), when in fact it has not.¹

5.1.3 Effects of AI explanations on human understanding

While one of the main goals of XAI is to enable users to understand the basis for the AI’s predictions (Adadi and Berrada 2018; Molnar 2025), empirical studies indicate that AI explanations do not necessarily have this effect. One study demonstrated that when users were given local feature contribution explanations (SHAP; see section 5.1.4 below) for predictions by an opaque model in a graphical interface, they over-estimated their own understanding of the AI’s overall logic (Chromik et al. 2021). Eiband et al. (2019) found that users perceived that they understand the workings of an AI better as long as they were given an explanation, even if the explanation did not reveal any information about these workings.

One limitation of previous studies is that they do not focus on how users’ comprehension process might be affected by the *specific form and content* of AI explanations. For example, Chromik et al. (2021) observe that some participants assumed monotonic feature effects, and judged the AI’s violations of this assumption as “illogical”, but do not engage in this finding or discuss general implications of it. The present chapter complements previous work regarding human understanding of AI explanations by explicitly analysing how particular kinds of AI explanations (enthymematic arguments) might be interpreted pragmatically.²

¹The concept is similar to the well-known *illusion of explanatory depth* (Rozenblit and Keil 2002), but focuses on the role of explanatory communication rather than explanatory depth/complexity.

²For a similar analysis in the context of AI-assisted decisions concerning employment support, see Berman (2024a).

5.1.4 Previous work on extracting arguments from ML models

As mentioned in section 2.5, previous work in XAI for ML can broadly be grouped into two subfields: (1) techniques for extracting explanatory information from arbitrary (including opaque) models using model-agnostic approaches, and (2) interpretable approaches to ML. Generally, neither of these subfields of XAI have been explicitly concerned with argumentative explanations or enthymemes. Exceptions include Liedeker et al. (2025) who propose a model-agnostic method for generating arguments based on feature attribution and Albini et al. (2021) who propose a method for extracting arguments from neural networks. Nevertheless, many techniques and approaches discussed in XAI research relate to argumentation, even if such a connection is not always articulated by authors. Hence, in what follows, the feasibility of extracting arguments for predictions from opaque and interpretable models will be discussed in relation to previous work, regardless of whether such work has been explicitly framed in relation to argumentation and enthymemes.

Black-box models and model-agnostic explanation methods Generally, arguments are difficult to extract for black-box models such as deep neural networks and random forests due to their complex inner workings. To some extent, “local” and “global” post-hoc explainability methods can be conceived to identify data and warrants respectively. Feature attribution methods such as LIME (Ribeiro, Singh and Guestrin 2016) and SHAP (Lundberg and Lee 2017) can be understood as data extractors, since they estimate to what extent features contribute positively and negatively to the prediction. Feature attributions can be expressed straightforwardly as statements about feature values (Forrest et al. 2018; Liedeker et al. 2025). For example, if LIME highlights BackPain (10-item scale) as the strongest positively contributing feature for a prediction concerning whether a patient will be satisfied with surgery, one can, in principle, construct the datum “The patient’s back pain level is 4” to explain the prediction.³ Note, however, that feature attributions do not convey what property of the feature value that makes it contribute positively or negatively to the prediction (e.g. whether a back pain of level 4 is deemed low or high).

Methods for generating counterfactual explanations, such as DiCE (Mothilal,

³Occasionally in previous work, feature attributions are expressed without indicating the feature value. The CXAI system TalkToModel (Slack et al. 2023) generates explanations such as: “For the model’s predictions on instance with id equal to 1234, the importance of the features have the following ranking, where 1 is the most important feature: 1: BackPain, 2: ...”. A similar approach is used by He, Aishwarya and Gadiraju (2025). In a real-world setting, the Swedish Public Employment Service explains ML-based decisions with reference to “the following factors that contributed to the assessment: your unemployment duration, your unemployment history, ...” (Berman 2024a).

Sharma and Tan 2020), can also be characterised as data extractors since they highlight features whose values can explain why a particular outcome is predicted rather than another one. For example, DiCE enables explanations such as: “The patient would be predicted satisfied with surgery if pain level was 3 instead of 4 and disability was 36 instead of 45.” However, datum-based explanations extracted using feature attribution or counterfactual explanation generation cannot be faithfully linked with warrants, as will be discussed below.

Data can also be identified by training a “surrogate” decision tree to mimic the behaviour of the actual model, and then explain specific predictions by conveying the relevant conditions for the prediction at hand (Sokol and Flach 2020; He, Aishwarya and Gadiraju 2025), e.g., that a patient is predicted to be satisfied with surgery since “the patient’s disability is below 43 and the patient’s back pain level below 4” (as discussed in the context of interpretable models below). However, since a surrogate model does not necessarily reflect the actual logic of the model, such explanations can be unfaithful.

In order to extract warrants—e.g. that patients with mild back pain are deemed more satisfied with surgery—various options can be conceived. However, due to model opacity, all of these options suffer from faithfulness issues. One option is to implicate warrants by conveying mean feature values together with feature attributions.⁴ However, warrants implicated by such information can unfaithfully reflect the actual model behaviour globally. For example, if a back pain level of 4 in relation to population mean of 5.5 is presented as contributing positively to the prediction that the patient will be satisfied with surgery, this naturally implicates that the model associates milder back pain with satisfaction with surgery. However, for an opaque model with potential non-monotonicities, there is no guarantee that this warrant generally holds. In fact, for neural networks and other models that are sensitive to small perturbations in input, it might not even hold for cases that humans perceive as highly similar (Szegegy et al. 2014; Ballet et al. 2019). Similarly, Nguyen, Schlötterer and Seifert (2023) generate counterfactual explanations in ways that linguistically trigger a warrant by indicating whether specific feature values are “too low/high”. For example, the fact that a patient is predicted unsatisfied rather than satisfied with surgery can be explained by stating that “back pain is too high”. However, warrants implicated in this manner are not necessarily faithful to the actual workings of the model, as argued above.

Another option is to use a global feature attribution method to convey information concerning overall model behaviour. For example, a partial dependence plot (PDP) (Friedman 2001) can show whether the relation between a feature and the predicted outcome is monotonic or more complex (Molnar 2025). However,

⁴Cinà et al. (2023) use such an approach to textually and graphically explain predictions concerning adverse outcomes after ICU discharge.

as far as the author is aware, there is no prior work on expressing outputs from PDPs or similar methods in natural language.

Warrants can potentially also be extracted by globally approximating model behaviour by training a more interpretable surrogate model with outputs from the target model as ground truth (Molnar 2025). For example, with a linear surrogate model, the monotonicity and effect size of relations between features and outcome can be generated using the method proposed in section 5.2, with the caveat that such a warrant would only reflect model behaviour approximately.

Another possibility which has been discussed is to generate an approximate linguistic summary of patterns in the training data. Kaczmarek-Majer et al. (2022) propose a method for accompanying local feature importance explanations (e.g. that BackPain is the most important feature for the prediction) with a global linguistic summary based on fuzzy quantified sentences of the type “Among patients with a high predicted probability of being satisfied, most of them have low back pain levels”. However, since summaries are approximate and not grounded in model behaviour, they can be unfaithful with respect to the model’s actual reasoning. (In the method proposed by Kaczmarek-Majer et al. (2022), linguistic summaries are qualified with their “degree of truth” on a scale from 0 to 1.)

Integration of external domain knowledge has also been proposed as a means to explain the outcome of feature importance attribution, i.e. why certain feature values influence a prediction (Feustel et al. 2024; Schindler et al. 2025). For example, if the user asks why loan duration influences decisions concerning credit approval, the system might reply: “Lenders will usually feel more comfortable lending you money for a shorter period because you’re more likely to be able to pay it back.” Such answers can be based on manually crafted knowledge base (Feustel et al. 2024) or on automated mining of relevant texts (Schindler et al. 2025). Argumentatively, external domain knowledge of this kind can be characterised as warrants, since it attempts to explain the link between data (feature importances) and claims (predictions). However, since external domain knowledge is not grounded in the model whose predictions are being explained, they can potentially lead users to form incorrect understandings of the workings of the model.

Interpretable models and model-specific explanation methods Unlike black-box models, intrinsically interpretable models make their predictions using reasoning steps that are, at least in principle, intuitive for humans to understand, and are therefore often considered favourable from the perspective of explainability (Rudin et al. 2022).⁵ Examples of statistical models that are typically characterised

⁵Some authors use the term “transparency” instead of “interpretability” (see, e.g., Lipton 2018; Barredo Arrieta et al. 2020).

as interpretable include linear models, decision trees, decision lists, and decision sets (Lakkaraju, Bach and Leskovec 2016; Sokol and Flach 2020; Rudin et al. 2022; Marques-Silva and Ignatiev 2023; Molnar 2025). There are also interpretable approaches in non-classical logic which can potentially be trained on data, such as fuzzy rule-based and possibilistic systems (Baaj 2022).

It should be emphasised that the distinction between interpretable and opaque models is not clear-cut (Rudin et al. 2022). For example, the extent to which a neural network is considered interpretable or opaque largely depends on its size (number of input/output neurons and hidden layers). In this sense, the distinction between model-agnostic and model-specific explanation methods is more appropriate; the former treat the explained model as a black box (regardless of its actual internal mechanisms), while the latter leverages knowledge about the workings of the model.

While a comprehensive discussion of the ability to extract arguments in model-specific ways is beyond the scope of this thesis, a brief comparison of three model-specific approaches can illuminate the choice of focus of the present chapter.

Even if *neural networks* are commonly conceived as opaque (Burrell 2016; T. Miller 2019; K. d. Fine Licht and J. d. Fine Licht 2020), knowledge about their inner workings can in principle be leveraged for argumentative explanations. Albin et al. (2021) propose Deep Argumentative Explanations (DAX), a method for treating (groups of) neurons as arguments by building a graph of support and attack relations on the basis of neural influence, enabling explanations such as “The patient is predicted to be satisfied with surgery since back pain level is 3, despite disability being 45.” However, as with feature attribution methods, DAX does not extract information concerning what property of the feature value that determines whether it supports or contradicts (attacks) the prediction.

Decision trees are typically considered interpretable since globally, the overall logic of a decision tree can easily be visualised, while locally, a decision can be explained by tracing it backwards through the conditions in the tree’s nodes. A prediction can be explained in natural language by stating the locally relevant conditions conjunctively (Maruf et al. 2023), argumentatively corresponding to data. For example, we can consider a model that has learned to predict outcome of surgery for patients with spinal stenosis according to the flow depicted in fig. 5.1; for this model, a prediction that the patient will be satisfied with surgery can be explained as: “The patient’s disability is below 43 and the patient’s age is below 73”. Alternatively, numerical conditions can be expressed in scalar terms (Alonso and Bugarín 2019; Stepin et al. 2020), e.g.: “The patient’s disability is *relatively low*”. Extracting warrants for decision trees is less straightforward due to the local nature of the conditions that constitute a decision tree’s logic. In

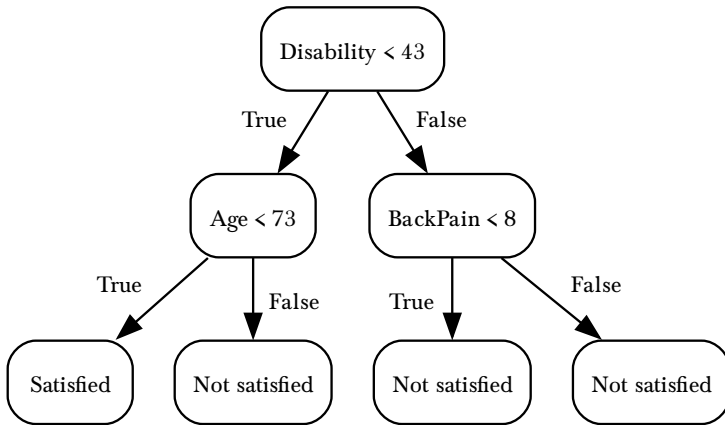


Figure 5.1: Example of a sparse decision tree for predicting outcome of surgery for patients with spinal stenosis.

principle, a warrant can be constructed for each path from root to leaf, e.g.: “The model associates low disability and young age with being satisfied with surgery”. However, since the number of paths grows exponentially with tree depth, this can lead to a very large number of warrants, each with undesirably narrow scope.

The method proposed here focuses on another class of models, namely *GLMs* (McCullagh and Nelder 1989). GLMs are commonly used in high-stakes domains such as healthcare, in part due to their interpretability (Pantanowitz et al. 2024). Traditionally, GLMs are explained “globally” by characterising the learned coefficients, e.g. in terms of monotonicity (positive or negative) and effect size, and “locally” by visualising how specific features contribute to the prediction, e.g. in an effect plot (Molnar 2025). The present work focuses on how the specific formal properties of GLMs can be leveraged for argumentative and enthymematic explanations, which, to the best of my knowledge, has not been done earlier. In contrast to black boxes and decision trees, GLMs force features to affect output independently of each other, without any interactions (assuming that no interaction terms are used). Furthermore, features affect output monotonically; for example, less back pain always increases the predicted probability that a patient will be satisfied with surgery. Due to these formal properties, warrants such as “The model associates milder back pain with a higher expected probability of being satisfied with surgery” can faithfully reflect the actual knowledge learned by the model. Furthermore, due to the independent treatment of features, each warrant has global scope. Warrants can also be triggered linguistically in datum

expressions by characterising numerical feature values as “relatively low/high” (see section 5.2).

5.2 Extracting arguments from GLMs

The proposed method for extracting arguments from GLMs enables predictions to be supported by data and/or warrants in ways that faithfully convey model behaviour, taking into account expected pragmatic inferences drawn by users. The method leverages specific mathematical properties of GLMs, namely lack of interactions between features, and monotonic relations between features and outcome. Specifically, the method enables claims concerning whether a particular outcome is expected to be higher or lower than for a reference instance, e.g. that there is a high probability that a particular patient will be satisfied with surgery compared to an average patient, to be supported by two kinds of arguments:

- **Datum:** A particular feature value is high or low compared to the reference instance, e.g., that the patient has relatively mild back pain.
- **Warrant:** A particular feature has a (positive or negative) monotonic relation to the expected outcome, e.g. that the model associates milder back pain with a higher expected probability of being satisfied with surgery.

Mathematically, datum-based arguments are extracted by identifying features that contribute either positively or negatively to the expected outcome (compared to the reference), while warrants are extracted by considering the sign of the model’s learned coefficients, as will be described in more detail in the subsequent sections.

5.2.1 Model specification

GLMs estimate an outcome on the basis of a linear combination of predictors (independent variables) and a link function that transforms the linear combination to an outcome:

$$\mathbb{E}[Y | X] = g^{-1} \left(\beta_0 + \sum_i X_i \beta_i \right) \quad (5.1)$$

where $\mathbb{E}[Y | X]$ is the expected value of outcome Y given the independent variables X , coefficients β_i (including the intercept/bias term β_0), and the (monotonic) link function g . For example, in the case of logistic regression, the expected value is a probability, and g is the logit function:

$$g(p) = \log\left(\frac{p}{1-p}\right)$$

Note also that since g is monotonic, the inverse link function g^{-1} is also monotonic.

5.2.2 Featurisation

The independent variables X are obtained as a numeric encoding of raw inputs \tilde{X} :

$$X = \phi(\tilde{X})$$

In principle, \tilde{X} can contain any types of features that can be encoded numerically, including not only scalar-valued (binary, ordinal, integer-valued and continuous) features, but also multinomial categorical features encoded with binary dummy variables. However, since the proposed method hinges on monotonic relations between features and outcome, it is primarily intended to be applied for scalar-valued features. (Handling of multinomial categorical features will be further discussed in section 5.4.2.)

5.2.3 Extracting data

A claim made on the basis of a GLM prediction can be said to be supported by feature values that contribute either positively or negatively to the estimated probability. For example, if the model predicts that a patient has a 0.85 probability of being satisfied with surgery, which is higher than average, and the patient’s relatively mild back pain contributes positively to the prediction, then the claim “The patient has a high probability of being satisfied with surgery” can be supported by the datum “The patient has relatively mild back pain”, using the method elaborated below. The same method can be used to extract data that contradict a particular claim, by focusing on features whose values contribute in the opposite direction.

Mathematically, eq. (5.I) can be reformulated with respect to the reference point R :

$$\mathbb{E}[Y | X] = g^{-1}\left(\beta_0 + \sum_i R_i \beta_i + \sum_i \beta_i (X_i - R_i)\right) \quad (5.II)$$

where $\beta_i (X_i - R_i)$ is the contribution for predictor i .

Data for integer-valued and continuous features (such as amount of back pain) can be expressed as statements concerning whether the value is relatively low or high (compared to the reference). Note, however, that for bounded numerical features, it may be adequate to express extreme values in ways that reflect their extremity, rather than their relation to the reference. For example, a value of 0 on a pain scale from 0 to 10 can be expressed as “The patient has no pain” rather than “The patient has relatively mild back pain”, since the latter might falsely implicate that the patient has some pain (Hirschberg 1985). Importantly, data expressed accordingly convey the property of a feature value that makes the expected outcome high or low. Data for binary features, e.g. whether the patient has other illnesses, can straightforwardly be expressed as statements concerning the feature value as such, e.g. “The patient has no other illnesses”.

Formally, we define $\text{DatumPhrase}(f, x, r)$ as a general function that returns a datum phrase:⁶

$$\text{DatumPhrase}(f, x, r) = \begin{cases} \text{FeatureLevelPhrase}(f, x, \text{RelValue}(x, r)) & \text{if } f_t = \text{num}, \\ \text{FeatureValuePhrase}(f, x) & \text{if } f_t = \text{bin} \end{cases} \quad (5.III)$$

where f contains information about the feature (including the feature type f_t which is `num` for numeric features or `bin` for binary features), $x = \tilde{X}_i$ is the raw feature value, and $r = \tilde{R}_i$ is the raw reference value. $\text{FeatureLevelPhrase}(f, x, l)$ is a grammar function that generates a feature-level phrase for feature f with value x and level l , e.g. “The patient has relatively mild back pain”. The grammar function $\text{FeatureValuePhrase}(f, x)$ returns a phrase stating a feature value, e.g. “The patient has other illnesses” for $f_{\text{name}} = \text{HasOtherIllness}$ and $x = 1$. The general function $\text{RelValue}(x, r)$ returns the relative level of feature value x in relation to a reference value r . Specifically,

$$\text{RelValue}(x, r) = \begin{cases} \text{high} & \text{if } x > r, \\ \text{low} & \text{otherwise} \end{cases} \quad (5.IV)$$

It should be noted that feature level characterisations explain why a feature value contributes positively or negatively to the expected outcome, *relative to a reference* (R). In other words, the choice of reference can influence both whether a specific feature value is considered low or high in the context, and whether the expected outcome is lower or higher than for the reference—even if the choice of reference as such does not influence the prediction. To illustrate this point with concrete examples, we can consider the feature `BackPain` (raw index i ,

⁶Monospaced font (e.g. `DatumPhrase`) is used for implementation-level entities.

encoded index j , negative coefficient) in relation to estimated success of surgery, considering two different reference patients:

- R^A : Average patient with same diagnosis, with $R_i^A \approx 6$.
- R^I : Patient with ideal circumstances for surgery, with $R_i^I = 1$.

Assuming a hypothetical patient with $\tilde{X}_i = 4$, the role of BackPain in relation to the claim will be expressed differently depending on choice of reference:

- For $R = R^A$, $\beta_j(X_j - R_j) > 0$ and $\text{RelValue}(\tilde{X}_i, \tilde{R}_i) = \mathbf{low}$. This can be expressed as: “The patient’s relatively *mild* back pain contributes *positively* to the estimated probability of successful outcome.”
- For $R = R^I$, $\beta_j(X_j - R_j) < 0$ and $\text{RelValue}(\tilde{X}_i, \tilde{R}_i) = \mathbf{high}$: This can be expressed as: “The patient’s relatively *severe* back pain contributes *negatively* to the estimated probability of successful outcome.”

As will be shown in section 5.3, this kind of context-specificity makes inferences that are naturally invited by arguments extracted in this manner faithful with respect to the actual reasoning of the model.

It can also be noted that the proposed approach for extracting data assumes a simple dichotomisation between low and high; this limitation will be discussed in section 5.5.

5.2.4 Extracting warrants

Warrants can be extracted on the basis of the sign of the coefficients. If $\beta_i > 0$ for a numerical feature, then higher values for predictor i contribute to a higher expected outcome, and vice versa. For example, if β_i for amount of back pain is negative, then less back pain contributes to a higher expected probability of being satisfied with surgery. For binary features, extracted warrants convey whether a particular value, e.g. having other illnesses, contributes to a higher or lower expected outcome.

An extracted warrant can be expressed with explicit propositional attitude, e.g. “The model associates milder back pain with a higher expected probability of being satisfied with surgery”. Such statements are inherently faithful with respect to how a GLM makes its predictions. Another option is to express warrants using generic statements, e.g. “Patients with less back pain are more likely to be satisfied with surgery”. However, this could potentially invite false inferences concerning actual matters of fact in the world. In cases of multi-collinearity, model coefficients may not reflect statistical correlations in the data (Lipton 2018; Molnar 2025). For

example, if, in reality, amount of back pain is strongly correlated with duration of back pain, and both correlate negatively with being satisfied with surgery, then one of the learned coefficients may end up positive. In such a scenario, the statement “The model associates a long duration of back pain with being satisfied with surgery”, expressed as a propositional attitude of the model, could be true (with respect to the actual properties of the model), while the generic statement “Patients with long duration of back pain are more likely to be satisfied with surgery” could simultaneously be false (with respect to the reality being modelled). Consequently, expressing extracted warrants with explicit propositional attitude is generally preferable from the perspective of faithfulness.⁷

By considering the magnitude of the coefficient and the absolute maximum slope of the inverse link function, warrants can also include information about effect size, e.g. “The model associates milder back pain with a higher expected probability of being satisfied with surgery. *The difference can be up to 24 percentage points.*”

5.3 Formal guarantees on correctness of invited inferences

An important property of the proposed method is that, under the assumptions regarding addressees’ pragmatic reasoning laid out in section 5.1.1, it can be shown that inferences that are naturally invited by enthymematically supporting predictions by either data or warrant are formally guaranteed to be correct. These guarantees apply to any numeric feature and concerns both inferences from data to warrant and vice versa, as elaborated below.

5.3.1 Inferences from datum-based enthymeme

If a feature-level datum phrase is communicated in support of a GLM-based claim, this can be assumed to invite inferences regarding both the feature-value range and the relation between the feature and the outcome at hand. For example, if a patient’s “relatively mild back pain” supports the claim that the patient is likely to be satisfied with surgery, this can be assumed to invite the following inferences:

- **Feature value:** The patient has less back pain than an average patient, i.e. $\bar{X}_i < \bar{R}_i$.

⁷Issues concerning the semantics of extracted arguments in the context of dialogue modelling are discussed in sections 6.5.6 and 6.5.7.

- **Feature–outcome relation:** The model associates mild back pain with a higher probability of being satisfied with surgery than severe back pain, i.e. $\mathbb{E}[Y | X]$ decreases monotonically with respect to X_j .

Importantly, the exemplified inferences are not just incidentally true in this particular case. Rather, their correctness follows formally from the mathematical properties of GLM and the argument extraction method. Specifically, the correctness of the feature-value inference follows from the condition for selecting the feature-level phrase “relatively mild back pain” ($\text{ReLValue}(\tilde{X}_i, \tilde{R}_i) = \text{low}$ iff $\tilde{X}_i < \tilde{R}_i$). The correctness of the inference regarding feature–outcome relation is guaranteed as follows. Presenting the relative feature value in support of the claim implicates that $\beta_j(X_j - R_j) > 0$, i.e. that the patient’s relatively mild back pain contributes positively to the estimated probability that the patient will be satisfied with surgery. Since $X_j - R_j < 0$ (because $X_j < R_j$), the product $\beta_j(X_j - R_j)$ can only be positive if $\beta_j < 0$, i.e. if $\mathbb{E}[Y | X]$ decreases monotonically with respect to X_j , as implicated.

5.3.2 Inferences from warrant-based enthymeme

Similarly, if the claim in section 5.3.1 is supported enthymematically using the extracted warrant, i.e. “The model associates milder back pain with a higher expected probability of being satisfied with surgery”, this invites the feature-level inference that the patient has relatively mild back pain, i.e. $\tilde{X}_i < \tilde{R}_i$. The correctness of this inference is guaranteed as follows. Presenting the warrant in support of the claim implicates that the patient’s relatively mild back pain contributes positively to the estimated probability that the patient will be satisfied with surgery, i.e., $\beta_j(X_j - R_j) > 0$. Since $\beta_j < 0$, this inequality can only hold if $X_j - R_j < 0$, i.e., $\tilde{X}_i < \tilde{R}_i$, as implicated.

5.3.3 Assumptions concerning expected monotonicity

The formal guarantees demonstrated above assume that AI users expect presupposed warrants to be monotonic. While this assumption is grounded in previous theories (see section 5.1.1), it is worth noting that the guarantees also hold in the more general case where users only expect monotonicity for a certain (typical) value range. For example, a user may infer that the model associates mild back pain with a higher probability of being satisfied with surgery than severe back pain, *but not necessarily for very mild back pain* (based on the intuition that if a patient has almost no pain at all, the patient might be less likely to experience an improvement after surgery). In this scenario, the user leaves the possibility open for an overall non-monotonic relation between back pain and satisfaction

with surgery. Importantly, such an interpretation is compatible with respect to how the AI actually reasons, since the perceived ambiguity concerning whether the relation is monotonic in a certain range does not contradict that the relation is in fact monotonic.

5.3.4 Relation to opaque models

It should be emphasised that correctness guarantees for naturally invited inferences cannot generally be provided when similar kinds of enthymematic explanations are conveyed for predictions by opaque models. To illustrate this point, we can consider a scenario where a random forest predicts that a patient will be satisfied with surgery. LIME or SHAP identifies BackPain as an important feature for the prediction, and the patient’s value for BackPain is below average ($\tilde{X}_i < \tilde{R}_i$). On this basis, the patient’s “relatively mild back pain” is communicated in support of the prediction. In this case, the warrant inference invited by the explanation—namely that the model associates milder back pain with a higher expected probability of being satisfied with surgery—is not necessarily true. For example, the random forest might assume an inverse relation between back pain and satisfaction with surgery for some combination of feature values.

5.4 Case study: Spinal surgery decision-support

To investigate the feasibility of applying the proposed method in a real-world use case, the approach has been implemented and preliminarily evaluated in the context of a web-based instrument used by spine clinics in Sweden. The purpose of the AI-based “Dialogue Support” tool is to assist doctors and patients during brief (approx. 20 minutes) medical consultations where decisions concerning choice of treatment (usually surgical or non-surgical treatment) are made. Using GLMs, the tool predicts two patient-reported outcomes and length of in-hospital stay of hypothetical surgery for patients with degenerative spinal disorders. The present study explores how the currently deployed instrument, which does not offer patient-specific explanations, can be modified and extended to meet doctors’ and patients’ clinical needs related to explainability.

The AI-based tool assists patients and doctors in their decision-making concerning treatment options for four different types of degenerative spinal disorders: disc herniation and spinal stenosis in the lumbar spine respectively, or chronic low back pain, as well as cervical radiculopathy (Fritzell, Mesterton and Hagg 2022). Based on sociodemographic information (age, gender, etc.) and other information provided by the patient in a questionnaire, the tool presents predictions for three

types of outcomes of a hypothetical surgery:

- **Satisfaction:** the probability of responding to the question “What is your attitude towards the result of the surgery?” with one of the first two options in the Likert scale *satisfied, hesitant, dissatisfied* one year after surgery.
- **Global assessment of pain (GA):** the probability of responding to the question “How is your pain today as compared to before the surgery?” with one of the first two options in the Likert scale *completely pain-free, much better, somewhat better, unchanged, worse* one year after surgery.
- **Length of stay:** the number of days of hospitalisation in connection with the surgery.

The predictions are based on three different types of GLMs trained on data from the national quality registry Swespine, with the following link functions:

- *logit* (i.e. logistic regression), for estimating satisfaction,
- *threshold function for the cumulative distribution function* (i.e. ordered probit), for estimating GA, or
- *rounding to non-negative integer* (linear (Ridge) regression adapted to counts), for predicting length of stay.

One model is used for each combination of diagnosis and task (type of outcome). Since there are 4 diagnoses and 3 tasks, a total of 12 models are used.

The existing version of the Dialogue Support tool explains its predictions “globally”, with information about features, sample size, etc. No case-specific (“local”) explanations are presented, while the version developed for the purposes of this study uses the proposed method to also provide local explanations.

5.4.1 Datasets

Historical patient data was obtained from the Swespine registry in the form of one dataset per diagnosis. The 4 datasets together encompass 37 features. Table 5.1 presents the feature types, with examples of features.

For the purposes of the study, features were selected to jointly optimise for performance and sparsity. This was done using backward elimination, with area under the ROC curve (AUC) as the performance metric for satisfaction and GA, and mean absolute error (MAE) for length of stay. Among the best-performing feature sets (performance not worse than the best performance minus a tolerance threshold), the feature set with the smallest number of features was selected.

Table 5.1: Feature types among the datasets used to train the predictive models. The reasoning behind the choice of encodings is described in section 5.4.2.

Feature type	Encoding	Example(s)
integer	standardisation	age, BMI
continuous	standardisation	quality of life (EQ-5D)
binary	binary	gender, has other illness
multinomial categorical	one-hot	clinic type (public, private, or university hospital)
high-cardinality ordinal	standardisation	pain levels (10-item scale)
low-cardinality ordinal	one-hot or standardisation	walking distance (5-item scale from 0–100 meters to more than 1 kilometer)

5.4.2 Feature encoding

Integer-valued features and continuous features were encoded numerically and standardised. For multinomial categorical features, one-hot encoding was used, thereby treating each value as a separate binary feature. It should be noted that with this treatment, warrants are not straightforwardly triggered. For example, if the fact that the patient is “treated at a university clinic” (rather than a public or private clinic) is presented as indicative of a negative outcome, a corresponding warrant cannot easily be identified. Specifically, the formulation does not convey whether the prediction would be more positive if treated in a public or private clinic. Since clinic type was the only multinomial feature among the most predictive features in the studied case (included in 7 of the 12 models), this limitation was considered acceptable.

As for ordinals, at least two approaches can be conceived: to treat them either as numeric or as multinomial categorical features. Numeric encoding is more favourable from the perspective of intelligibility, since the order of values enables scalar warrant triggers. However, a categorical encoding may yield better performance. To balance these considerations, high-cardinality ordinals (> 5 values) were encoded numerically, since it was deemed highly unintuitive to treat each value on a 10-item pain scale as its own category, while for low-cardinality ordinals, a data-driven approach was employed to jointly optimise for intelligibility and performance. Among the best-performing encoding candidates, the feature set with the largest amount of numeric encodings was selected. The result of these data-driven featurisation choices was that only one of the 33 low-cardinality

ordinal features (across the 12 models) was encoded categorically, while the rest were encoded numerically.⁸

The proposed strategy for choosing feature encodings is summarised in table 5.1.

5.4.3 Interface design

Data are presented in a waterfall chart, where the estimated outcome is visualised in terms of the outcome for an average patient, and the cumulative effect of data (see fig. 5.2), grouped into positive and negative and ordered by decreasing importance (effect size). For example, the fact that the probability of a successful outcome is lower than for an average patient can primarily be explained with reference to the fact that the patient is treated at a university hospital and has a relatively long duration of leg pain. A maximum of three positive and negative factors respectively is shown by default; additional data can be revealed by clicking “Show more”. Conceptually, the presented information explains both why a specific absolute probability (e.g. 70%) is predicted, and why the predicted probability is lower/higher than for an average patient (e.g. 70% for the patient at hand, rather than 61% for an average patient).

Outcomes are visualised along a probability scale (for satisfaction and GA) or integer scale (for length of stay), while data (feature contributions) are visualised without any explicit scale. In other words, mathematically, the waterfall chart informally conveys:⁹

$$\mathbb{E}[Y | R] + \sum \beta_i(X_i - R_i) \approx \mathbb{E}[Y | X] \quad (5.V)$$

Warrants conveying information about the statistical patterns learned by the model, e.g. that lower disability is associated with a higher estimated probability of a successful outcome, are presented inside a widget titled “More information”. The widget also contains information about the sample size and training data, which argumentatively can be said to back the warrants. The content of the widget is collapsed by default, but can be expanded as needed.

5.4.4 Choice of reference point

As elaborated above, data are visually and linguistically presented in relation to a reference point. In the proposed prototype, the reference is chosen as the mean

⁸Predictive performance after data-driven featurisation choices was AUC 0.65–0.69 for satisfaction, AUC 0.62–0.69 for GA, and mean absolute error 0.27–1.06 for length of stay.

⁹A completely faithful visualisation would need to account for non-linearity of the link function. This degree of faithfulness was not deemed motivated for the purposes at hand.



Figure 5.2: Screenshot of an English version of the prototype, containing extracted argumentative explanations for GLM-based predictions regarding outcome of spine surgery, with a hypothetical patient profile.

historical patient with the same diagnosis ($R = \bar{X}$), although such a notion may be perceived as abstract. The main reason for not settling with the intercept/bias as reference is that this would introduce an undesired bias for binary features; e.g. the gender encoded as 0 (in this case being male) would never be highlighted as a factor.

The frequently observed comparison with expectation/norm in the medical corpus study (see chapter 3) may suggest using “healthy patient” as a reference in medical contexts where it can be clearly established *a priori* what constitutes a “healthy” feature value. Allowing reference to be chosen interactively may also be an option.

5.4.5 Generation of data and warrants

The proposed method for generating linguistic explanations for GLM predictions consists of general (domain- and language-independent) functions for generating data and warrants, based on the approach presented in section 5.2, which depend on a domain- and language-specific grammar containing functions for producing linguistic surface realisations.

Generation of data

For the grammar functions `FeatureLevelPhrase` and `FeatureValuePhrase`, the present work uses simple templates and mappings (see fig. 5.3). Given that the users can be expected to know that the displayed information concerns the patient at hand, and in order to avoid excess repetitions, data are formulated non-sententially (“No other illnesses”) rather than sententially (“The patient has no other illness”).

The 10 most frequently occurring generated factual phrases for historical patients (i.e. instances in the datasets) are shown in table 5.2 (positive factors) and table 5.3 (negative factors).

Generation of warrants

Warrants conveying correlations learned by the models are expressed in ways that communicate both effect size and, when relevant, polarity. For example, the warrant for a numeric feature can be formulated as: “The less back pain, the higher the calculated probability of a successful outcome. The difference can be up to 24 percentage points.” Effect sizes are calculated with respect to the magnitude of the coefficient and the absolute maximum slope of the inverse link

$$\begin{aligned}
& \text{FeatureLevelPhrase}(f, x, l) = \\
& \left\{ \begin{array}{l} \text{if } f_{\text{name}} = \text{BackPain} : \left\{ \begin{array}{ll} \text{“No back pain”} & \text{if } x = 0, \\ \text{“Relatively mild back pain”} & \text{if } x > 0 \wedge l = \text{low}, \\ \text{“Relatively severe back pain”} & \text{if } x > 0 \wedge l = \text{high} \end{array} \right. \\ \text{if } f_{\text{name}} = \dots \end{array} \right. \\
& \\
& \text{FeatureValuePhrase}(f, x) = \\
& \left\{ \begin{array}{l} \text{if } f_{\text{name}} = \text{HasOtherIllness} : \left\{ \begin{array}{ll} \text{“No other illnesses”} & \text{if } x = 0, \\ \text{“Has other illnesses”} & \text{if } x = 1 \end{array} \right. \\ \text{if } f_{\text{name}} = \dots \end{array} \right.
\end{aligned}$$

Figure 5.3: Examples of domain- and language-specific linguistic mappings in the domain of spinal disorders.

function.¹⁰ Generated warrants for a particular diagnosis and outcome are shown in table 5.4.

5.4.6 Evaluation

Following previous work on participatory design (Sanders and Stappers 2008), an early prototype of the proposed solution for argumentatively explaining predictions concerning outcome of hypothetical surgery was evaluated through a design workshop with orthopaedic surgeons. In the invitation, potential participants who all had experience of using the existing Dialogue Support tool were informed that they would test a new alternative interface to the tool. 3 out of 5 invited surgeons participated in the workshop on site, while one tested the prototype individually and then gave written feedback via email.¹¹

In the first part of the workshop, participants individually accessed the prototype, where a randomised patient profile was shown.¹² They were then asked to imagine having a dialogue with a patient who wants to know why the computer estimates an X% probability of successful outcome, and to answer the patient’s

¹⁰ $|\max((g^{-1})'(x))|$ is $\frac{1}{4}$ for logistic regression, $\frac{1}{\sqrt{2\pi}}$ for ordered probit, and 1 for linear regression.

¹¹ One of the research collaborators participated in the workshop in the role of orthopaedic surgeon. This collaborator had not been involved in the development of the prototype or the organisation of the workshop.

¹² The randomisation was done individually for each participant. Feature values were uniformly sampled from pre-defined ranges.

Table 5.2: The 10 most frequently occurring generated positive factual phrases for historical patients, aggregated across diagnoses and outcomes. “Occurring” here means that the datum is included among the top three positive factors shown in the interface. “Positive” refers to higher estimated probability of satisfaction or successful outcome, or shorter duration of stay. Frequencies are relative to the total number of patients.

Data presented as positive	Freq.
Treated in private clinic	0.46
No other illnesses	0.35
No previous spine surgery	0.32
Relatively low disability	0.18
Relatively few operated levels	0.17
Relatively short duration of back pain	0.15
Has university education	0.13
Relatively mild back pain	0.13
Can walk relatively far in normal pace	0.12
No university education	0.12

question very briefly. Participants noted their answers and were then asked to voluntarily read them aloud.

In the second part, participants were asked: “Is there anything in the explanations that can be improved?”¹³ The discussion was moderated by the organiser of the workshop (the author).

The version of the prototype tested by participants only supported one diagnosis (disc herniation) and two outcomes (satisfaction and GA). Instructions were conveyed to participants verbally and via a beamer presentation. Feedback was recorded in notes and later organised according to themes.

Results for part I

One of the participants volunteered to share the following explanation that he had prepared in response to the hypothetical patient’s question (own translation from Swedish):

There is a 76% probability of being satisfied with surgery which is 10% worse than an average pat[ient] who is operated on for disc herniation. The reason that it looks this way is that you have had back pain for

¹³Participants were also asked if there is anything else with the alternative interface that can be improved. Results from this part of the workshop are not reported here.

Table 5.3: The 10 most frequently occurring generated negative factual phrases for historical patients, aggregated across diagnoses and outcomes. “Occurring” here means that the phrase is included among the top three negative factors shown in the interface. “Negative” refers to lower estimated probability of satisfaction or successful outcome, or longer duration of stay. Frequencies are relative to the total number of patients.

Data presented as negative	Freq.
Relatively high disability	0.19
Treated in public clinic	0.18
Relatively severe back pain	0.18
Has other illnesses	0.16
No university education	0.16
Relatively long duration of back pain	0.15
Relatively short height	0.13
Relatively many operated levels	0.13
Relatively old age	0.13
Relatively long duration of leg pain	0.10

a long time and furthermore xxx¹⁴ diseases. Furthermore, your age gives you a statistically somewhat lower probability of a successful result. Since you are a pensioner [?] also looks like a factor that decreases the probability of being satisfied.

Overall, the answer indicates that this participant was able to interpret the explanatory information in the tool and produce a linguistic explanation in his own words, based on the provided information. Importantly, the participant’s explanation is consistent with the information in the tool, and thereby also with the actual reasoning logic of the model.

On a more detailed level, it is worth noting that although the question from the hypothetical patient concerned the *absolute* estimated probability (“why the computer estimates an X% probability of successful outcome”), the doctor chooses to explain the estimated probability *relative to an average patient*. It is unclear if the doctor focuses on this explanandum based on an interpretation that a 76% probability of satisfaction is low, or if design of the tool anchors such a framing (cf. Tversky and Kahneman 1974).

When the doctor expresses the specific causes of the prediction, the adverb “somewhat” is used in one instance to indicate relative importance, presumably

¹⁴The participant presumably noted “xxx” since the tool only stated that the patient has other diseases, without mentioning which ones.

Table 5.4: Generated warrants conveying correlations for spinal stenosis and pain assessment (GA). Warrants are shown to users when they click “More information” in the interface.

Feature	Warrant
Disability	The lower the disability, the higher the calculated probability of a successful outcome. The difference can be up to 33 percentage points.
Leg pain duration	The shorter the leg pain duration, the higher the calculated probability of a successful outcome. The difference can be up to 19 percentage points.
Previous spine surgery	Patients who have not undergone previous spine surgery are calculated to have a higher probability of a successful outcome. The difference can be up to 14 percentage points.
Back pain	The less back pain, the higher the calculated probability of a successful outcome. The difference can be up to 13 percentage points.
Back pain duration	The shorter the back pain duration, the higher the calculated probability of a successful outcome. The difference can be up to 11 percentage points.
Comorbidity	Patients with no other illnesses are calculated to have a higher probability of a successful outcome. The difference can be up to 8 percentage points.
Unemployment	Patients with employment are calculated to have a higher probability of a successful outcome. The difference can be up to 8 percentage points.
Type of clinic	Patients treated at private clinics are calculated to have the highest probability of a successful outcome. Patients treated at university hospitals are calculated to have the lowest probability of a successful outcome. The difference can be up to 7 percentage points.

reflecting a small effect size in the tool. In another instance, epistemic distancing (“*looks like* a factor that decreases the probability”) is used, presumably on a similar basis. These observations demonstrate some of the linguistic design choices through which an explainer can frame quantitatively presented information.

It is also worth noting that numerical data are both expressed in scalar terms (“you have had back pain for a *long* time”), presumably mirroring the scalar datum expression in the tool, and in one case without any numerical or scalar characterisation at all (“your age gives you...”). In the latter case, the omission of a scalar characterisation is presumably caused by an assumption that the patient already knows whether his/her age is considered low or high in the context. This can be contrasted with omission of scalar characterisation caused by model opacity, as discussed in section 5.1.4.

Results for part II

Two participants expressed difficulties regarding how to explain the prediction to the hypothetical patient, since they were unsure how to explain that the AI’s predictions were based on historical data. One of them asked if it would be adequate to tell the patient that the model makes its prediction “based on a group of operated patients who have responded similarly to you in the questionnaire”. The question might potentially suggest that these participants had an intuition concerning the AI that aligns better with some kind of case-based reasoning (CBR) than with GLMs. In response to the raised issue, information was added in the tool to clarify that predictions are made on the basis of the entire available sample of patients with the diagnosis at hand.

One participant commented that the textual explanations for the model and the specific outcomes were good and informative. Three suggestions related to global explanations (warrants) were raised. First, it was recommended to use positive instead of negative wording; e.g. substituting expressions like “The *older* the age, the *lower* the...” with “The *younger* the age, the *higher* the...”. Second, the wording “is estimated” was advised to be replaced with alternatives such as “is calculated”, “results have previously shown”, or “based on previous patients’ results of surgery”. Third, it was suggested that features should be sorted by descending effect size. All these suggestions were subsequently accommodated.

A question was raised as to whether the correlation between age and satisfaction is in reality “a curve rather than linear”. The question highlights a discrepancy between the participant’s domain knowledge and monotonicity assumptions built into the model. The fact that this discrepancy surfaced can be seen as a positive finding, in the sense that the tool’s explanation enabled the user to form a correct mental model of how the AI reasons (and to contrast this model with his/her own

reasoning).¹⁵

One participant observed that some factors were missing from the global explanations; this was later attributed to a bug, which has since been resolved.

In summary, the design workshop resulted in generally positive findings regarding the proposed linguistic explanations, as well as suggested improvements that have been accommodated into the prototype.

Discussion of results

While the observations in the design workshop were generally favourable, they also revealed a discrepancy between one of the participant’s expert-based assumptions and monotonicity constraints built into the model. Although the fact that the discrepancy surfaced can be seen as a positive finding (as discussed above), it also suggests that a strict focus on monotonic models can be undesirably rigid (as will be elaborated in section 5.6).

The observation that two workshop participants seemed inclined to explain model predictions in terms of case-based reasoning (CBR)—although uncertain if this would be accurate—raises a more general question concerning whether predictions by AI models based on CBR, e.g. k-nearest neighbours (Cover and Hart 1967), might be easier for humans to comprehend than, e.g., predictions from GLMs. It has previously been argued that CBR models might be appealing from the perspective of interpretability, since they better emulate how humans solve problems (Rudin et al. 2022). It has also been argued that explanations based on specific experience may be more acceptable to users than those based on general rules (McSherry 2001). However, from a linguistic perspective, explainability for CBR does not seem evidently aligned with how humans support and challenge judgements in linguistic interaction. In the non-AI-assisted human–human explanatory dialogues analysed in chapter 3, interlocutors do not back claims with explicit reference to similar cases. Instead, they frequently use scalar expressions such as “*low oxygen update*”, suggesting an overall conceived relationship between, e.g., oxygen saturation and choice of action. Presumably, if the interlocutors intended to convey some kind of case-based reasoning in support of their claims, they would have referred to “previous experiences of patients with an oxygen uptake of around 93%” or something along those lines. That said, the absence of such strategies in the relatively small corpora analysed in this thesis does not preclude that case-based explanations are *sometimes* used in human–human interaction. To this end, the potential usefulness of case-based linguistic AI explanations cannot be ruled out merely on the basis of the present study.

¹⁵Fitting a quadratic logistic regression model on the data indeed revealed a slightly curvilinear shape (with a peak around 35 years), but with a very small decrease in satisfaction for lower ages.

5.5 Limitations

The proposed approach for extracting data for numerical features assumes a simple dichotomisation between low and high compared to a reference point. An evident drawback of this assumption is that it fails to account for more fine-grained characterisations occasionally observed in the empirical material, such as “*pretty* dark (urine)” (excerpt 3.7), “*very* small (nares)” (excerpt 3.12) and “you *slightly* prefer non-danceable music” (excerpt 4.24).¹⁶

While it would be technically straightforward to map a feature value onto a gradable expression, e.g. by considering the specific value in relation to the statistical distribution of values for that feature (cf. Kaczmarek-Majer et al. 2022; Sokol and Flach 2024), this would not necessarily reflect the relative importance of the feature in relation to a GLM prediction, since feature contributions also depend on model coefficients and choice of reference point. For example, if feature A is somewhat above median while feature B is in the top decile, the relative contribution of feature A might still be larger than for feature B, if A has a significantly larger coefficient than B; consequently, explaining a prediction by characterising feature A as “somewhat high” and B as “very high” might undesirably invite false inferences concerning the relative importances of the features.¹⁷ While the proposed approach avoids this problem by adopting a simple dichotomisation, in future work it would be interesting to explore how more fine-grained gradable expressions can be obtained in ways that faithfully reflect the reasoning of GLMs.

5.6 Conclusions

Informed by linguistic analyses of human–human explanatory dialogues, this chapter has proposed a method for generating argumentative explanations for predictions from GLMs. The proposed method concisely explains GLM predictions (claims) with salient case-specific facts (data) and statistical patterns (warrants) and enables predictions to be explained enthymematically, e.g., with only data. Importantly, inferences that are naturally invited by such enthymemes are formally guaranteed to reflect the actual behaviour of the model. These guarantees hinge on certain interpretable properties of GLMs, namely independent treatment of features, and monotonic relations between features and outcome. Since these properties typically do not apply to black-box models, corresponding enthymemes

¹⁶Also cf. the use of the adverb “somewhat” by the workshop participant in section 5.4.6.

¹⁷Dependencies between judgements constitute another challenge with a more fine-grained characterisation of feature values, as will be discussed in section 6.3.6.

for black-box predictions can easily give rise to inferences that do *not* reflect model behaviour, potentially creating an illusion of explanatory grounding and causing users to form inaccurate mental models.

Arguably, the simplicity of the linguistic explanations obtained with the proposed method (e.g. that a patient will probably be satisfied with surgery because “the patient has relatively mild back pain”) reflects a desirable alignment with how humans typically handle explanations in linguistic interaction. Notably, this simplicity is not obtained at the cost of reduced faithfulness, unlike with popular post-hoc explanation methods for black-box models (such as LIME and SHAP).

While the proposed method applies to a class of models (GLMs) with relatively wide adoption (Maruf et al. 2023), its model-specificity can nevertheless be conceived as a limitation, for several reasons. From the perspective of pragmatic theory, as mentioned in section 5.1.1, human interlocutors may occasionally assume non-monotonic relationships between predicates (e.g., that walking distance can be either too short and too long), even when they explain their judgements in ways that seem to indicate a monotonic relationship (e.g., that a certain path is too short). If AI systems should ideally emulate human behaviour, it might therefore be relevant to support enthymematic explanations for predictions based on non-monotonic (e.g. quadratic) relations between features and outcome, potentially corresponding to a combination of two inversely related topoi (Ducrot 1988). Furthermore, even if humans find monotonic relations easier to process than more complicated functions (Busemeyer et al. 2013; Lucas et al. 2015), in some domains and for some features, (certain kinds of) non-monotonicity can be more intuitive (Rudin et al. 2022), as observed in the present evaluation. In addition, since non-monotonic functions are more general than monotonic ones, they can potentially enable more accurate AI predictions. In future work, it would therefore be useful to investigate how the method proposed here can be extended to additive models with non-monotonic relations between feature and outcome, e.g. generalised additive models (GAMs) (Hastie and Tibshirani 1986) or shape constrained additive models (SCAMs) (Pya and Wood 2015).

The proposed method for generating explanations for predictions from GLMs has been preliminarily evaluated in the context of an AI-based decision-support concerning degenerative spinal disorders. In a design workshop, 4 orthopaedic surgeons tested and gave feedback on a statistical tool which explains three different kinds of predictions regarding outcomes of hypothetical surgery using the proposed method. While the overall findings from the design workshop were largely positive, more work will be needed to study how AI explanations obtained with the proposed method are interpreted and used by human decision-makers in real-world scenarios. It would also be useful to study human communicative strategies when the proposed kind of AI explanations assist joint human decisions,

compared to similar situations with unexplained AI assistance or no AI assistance at all. These issues are addressed in an ongoing follow-up study.

5.7 Ethical considerations

Approval from the Swedish Ethical Review Authority was obtained (case number 2024-00839-01) for handling (de-identified) patient data from Swespine's registry.

CHAPTER 6

Modelling explanatory dialogue management

The studies of human–human explanatory dialogues in chapters 3 and 4 identified various communicative strategies used by human interlocutors when addressing the epistemic basis of complex judgements. One of the main findings was that interlocutors consistently explain judgements *enthymematically*, i.e., using arguments with implicit premises. It was also observed that explainers use certain linguistic *triggers* that seem to help explainees identify omitted premises. Beyond this observed regularity in the empirical material, the studies also revealed substantial variability concerning several aspects of explanatory dialogue behaviour, suggesting that both explanatory content and modes of interactive coordination of explanation delivery are *selected in context*. For example, interlocutors decide whether to explain judgements on demand or unpromptedly. Furthermore, different types of evidence are conveyed in support of judgements, including specific circumstances as well as general patterns (corresponding to Toulmin (2003)'s *data* and *warrants*). When claims are based on a chain of inferences, the explainer can select which inferential link to highlight. With respect to complex explanantia, multiple pieces of evidence can be delivered conjunctively in a single turn, or incrementally by jointly coordinating information delivery using feedback, backchannels and continuation markers.

The present chapter develops and discusses a formal model of dialogue management which is empirically grounded in the observed human explanatory strategies and normatively aligned with principles of transparency and faithfulness. In terms of empirical grounding, the model accounts for the regularities observed in the observed human dialogues, while simultaneously leaving room for the variabilities. Normatively, the model highlights aspects of explanatory interaction which enable explainees to form an accurate understanding of the epistemic basis of the explainer's judgements. The over-arching purpose of the research is both to seek a theoretical understanding of the high-level cognitive mechanisms used by agents engaging in dialogue concerning epistemic bases of complex judgements, and to enable development of explanatory dialogue systems with potential practical applications for human–AI interaction.

In relation to previous work, the dialogue model presented in this chapter extends prior work on general (non-ML-focused) argumentative dialogue systems by showing how claims corresponding to ML-based predictions can be challenged

and supported in dialogue. It also extends previous work on conversational interfaces to ML models by highlighting the role of warrants in human explanatory dialogue and showing how warrants that are faithful to the workings of the ML model can be elicited in human–computer dialogue.

The remainder of the chapter is organised as follows. Section 6.1 situates the present chapter in relation to previous work on explanatory and argumentative dialogue. The main contribution of the chapter is contained within sections 6.2 and 6.3 which discuss how the dialogue behaviours observed in dialogues in chapters 3 and 4 can be modelled formally and implemented computationally in a novel framework for explanatory dialogue management called BKOS.¹ Finally, section 6.5 discusses limitations of the proposed framework and how they might be addressed, while conclusions and potential future research directions are presented in section 6.6.

6.1 Related work

The subsections below position the present work in relation to prior research on modelling explanatory and argumentative dialogue, and introduce the notion of information-state based dialogue management, on which the proposed framework is based.

6.1.1 Modelling and implementing explanatory and argumentative dialogue

Some early explanatory dialogue systems were equipped with the ability to trace their own reasoning steps and recount such traces when a user asks the system to explain its outputs or behaviours. Examples includes SHRDLU (Winograd 1971), MYCIN (Shortliffe 1976; Buchanan and Shortliffe 1984) and Moore and Paris (1993)’s text planner for advisory dialogues. DAISY (Wahde and Virgolin 2023) can be seen as a recent incarnation in the same tradition.

The proposed dialogue model is not based on trace-based explanations and instead relates more closely to the tradition of modelling argumentative dialogue. Bench-Capon (1998) specifies and implements a “Toulmin Dialogue Game” which accounts for argumentative moves such as making a claim, asking a why-question and supplying a warrant. In a similar vein, Prakken (2005) formalises a class of dialogical argumentation games where each dialogue move attacks or surrenders to the other interlocutor’s previous move, while Black and Hunter (2008) formalise

¹The name BKOS is an amalgamation of the word *because* and Ginzburg (2012)’s term KoS which relates to conversation-oriented semantics.

inquiry dialogues involving enthymemes. Bex and Walton (2016) argue that explanations and arguments are different forms of reasoning and propose a unified dialogue model for both explanations and arguments. Hosseini (2016) formalises how interlocutors utilise beliefs about each other’s mental states to construct enthymemes whose implicit premises can easily be reconstructed by the receiver. In an effort to bridge logic-based argumentation models with real-world human communicative strategies, Xydis et al. (2020) propose a dialogue model which accounts for recovery from misunderstandings that may arise due to the use of enthymemes. Breitholtz (2020) presents a formal account of how claims can be justified enthymematically in dialogue, corresponding to backing of claims with data. Argumentative dialogue systems have been discussed in relation to a broad range of tasks and applications, including in the context of human–robot interaction (Sklar and Azhar 2018), wellness consultation (Sassoon et al. 2019), medication choices (Shaheen, Toniolo and Bowles 2020) and medical treatment (Castagna, Garton et al. 2023). For a review of argumentative dialogue systems, see Castagna, Kökciyan et al. (2024).

Generally, the prior work mentioned above does not target ML-based predictions. One exception is Sklar and Azhar (2018) who propose to unify computational argumentation and XAI by allowing arguments to be learned from data. However, the authors do not elaborate this proposition in detail.²

In the more specific context of explanations for ML predictions, several authors have proposed dialogue interfaces between human users and arbitrary ML models explained with the kinds of black-box explainability methods discussed in section 5.1.4 (Sokol and Flach 2020; Feldhus, Ravichandran and Möller 2022; Nguyen, Schlötterer and Seifert 2024; Wijekoon et al. 2024; Mindlin, Robrecht et al. 2024; Schindler et al. 2025; Mindlin, Booshehri and Cimiano 2026); for reviews, see Mindlin, Beer et al. (2025) and Berman and Larsson (2025). However, all of these approaches suffer from the problems discussed in section 5.1.4, namely that potential warrants implicated by generated explanations do not necessarily reflect the actual workings of the predictive model.

As for systems targeting interpretable models, Sokol and Flach (2018) present a dialogue system capable of explaining predictions by decision trees by interactively enumerating all counterfactual scenarios under which a different decision is obtained, starting with scenarios involving minimal feature changes. For example, if the user asks why a loan application was declined, the system may respond: “Had you earned more than €1000, instead of €750, it would be accepted”; the system can then provide additional scenarios when demanded by the user (Sokol

²As discussed in section 5.1.4, some previous work discusses argumentation extraction from ML models. However, prior research in this area does not typically target conversational modes of interaction.

and Flach 2020). The authors also discuss the possibility to provide decision rules (Sokol and Flach 2020), corresponding to warrants with narrow scope (see section 5.1.4). Braines, Preece and Harborne (2018) propose a conversational interface to a system consisting of high-level rules operating on outputs from ML-based sub-components. The interface can explain why a traffic situation is classified as congested or not congested as an implication from datum to claim. For example, a prediction that a particular situation is not congested can be explained with the utterance: “Live CCTV shows 12 moving objects at 80% of the speed limit which indicates it is not congested” (i.e. a statement of the form “*datum* indicates *claim*”). The interface can also explain ambivalent (uncertain) classifications using inconsistent evidence, e.g. “Live CCTV shows 2 moving objects at 80% of the speed limit which indicates it is not congested. But the live CCTV image is classified as congested. These outcomes are inconsistent.” However, the authors do not discuss the prospect of eliciting the system’s rules (warrants), e.g. that if objects are moving at 75% of the speed limit, or above, then the road is not congested, otherwise it is congested. Diana C. Hernandez-Bocanegra and Ziegler (2023) describe a system which can explain hotel recommendations based on an explicit factor model using data and backing. For example, if the user asks why a particular hotel has the highest rating, the system might reply: “Because of the positive comments reported regarding the aspects that matter most to you: 86% about location, and 85% about price.” Argumentatively, this answer highlights two pieces of data (that the hotel has 86% positive comments concerning location, and 85% positive comments concerning price); however, elicitation of warrants is not discussed. Bordt et al. (2024) study the prospect of using an LLM fed with a textual representation of an interpretable model as a conversational interface between user and model, and report positive findings concerning the ability of both GPT-3.5 and GPT-4 to analyse and describe the workings of the model. However, conversational strategies (such as argumentation and enthymemes) were not studied. In relation to faithfulness, the authors report that explanations provided by LLMs were found to be grounded in the actual workings of the model. However, their sample was very small, and a recent review of retrieval-augmentation for LLMs has concluded that confabulations remain a concerning issue (W. Zhang and J. Zhang 2025).

While some of the approaches described above are inspired or informed by human explanatory dialogue strategies, at least on an abstract level, the framework proposed here uses a bottom-up approach which is explicitly grounded in analyses of human explanatory dialogues. Specifically, the framework aims to model human-computer interactions which are distilled systematically from human-human dialogues (see sections 3.3 and 4.4). Furthermore, in contrast to previous work on argumentative/enthymematic dialogue, the present work shows how such

dialogue can be applied to statistical inferences based on ML models. Unlike previous approaches to conversational explainability for ML models, the proposed model handles warrants and enables explanations that are inherently faithful with respect to the workings of the model. Furthermore, a recent assessment of three CXAI systems, including a previous version of BKOS, found that only BKOS handled polar (yes-no) questions, nested explanations (explanations of previous explanations), and presupposition violations (Berman and Larsson 2025). An inability to handle presupposition violations was also observed among all 10 CXAI systems studied by Mindlin, Beer et al. (2025).

6.1.2 Modelling dialogue management using information states

The proposed approach for modelling explanatory dialogue builds on previous work on information-state based dialogue management (Larsson 2002; Ginzburg 2012; Maraev, Bernardy and Ginzburg 2020). Larsson (2002) conceives dialogue management as a process through which the *dialogue moves* performed by interlocutors, e.g., asking a question or providing an answer, trigger updates of individual agents' *information state* (IS), which represents the agent's view of the current state of the dialogue. An IS (or *dialogue gameboard*) may, in principle, contain any kind of information, including plans about what to say next, interpretations of what the other interlocutor has said, or beliefs about the world.

Information state updates (ISU) have been formally modelled in various different ways, including as rules embedded in Prolog (Larsson and Traum 2000; Larsson, Berman et al. 2000; Larsson 2002), as records or functions in type theory with records (Ginzburg 2012; Breitholtz 2020; Cooper 2023), and as TypeScript functions (Maraev, Berman and Larsson 2025). The proposed approach to ISU extends the work by Maraev, Bernardy and Ginzburg (2020), who model updates using a framework based on linear logic (implemented in Haskell), by enabling antecedents of linear implications to access knowledge and relations encoded in Prolog.

At the core of the modelling approach presented here is the notion that associations learned by statistical models can be captured by a particular kind of formal proposition, namely a *support* proposition. This kind of representation expresses how antecedents support consequents, e.g. the a person's mild back pain supports the assessment that the person will be satisfied with surgery. Support propositions are closely related to topoi (Ducrot 1988; Breitholtz 2020), warrants (Toulmin 2003), and defeasible rules in formal argumentation frameworks such as ASPIC+ (Modgil and Prakken 2014). Importantly for the present work, support propositions can be extracted from an interpretable model using the method

proposed in section 5.2.4 and enable the epistemic bases of predictions to be interrogated and explained by formally operating on antecedents, consequents and support relations between them, leveraging native Prolog mechanisms such as proof search and unification (Bratko 1986) for the purposes of dialogue modelling.

Another important principle behind the proposed modelling approach is that in situations where several valid courses of action can be identified, the policy for selecting an action is *external to the dialogue model* as such. In this sense, the proposed model is abstract and needs to be instantiated with concrete policies. This way, observed variability in the empirical material can be accounted for without making unsubstantiated assumptions concerning the causes or mechanisms behind the variability.

6.2 Preliminaries

Information-state updates are here conceived as linear implications (Maraev, Bernardy and Ginzburg 2020) and implemented in a framework based on SWI Prolog (Wielemaker et al. 2012). Specifically, the Prolog term $N :: A \text{ -* } C$ denotes an update rule named N (arbitrary term) defined as the linear implication $A \multimap C$. Update rules are processed using a simple control algorithm consisting of a loop which tests rules sequentially in their specified order (see algorithm 1 in appendix C).

An antecedent (left-hand side of linear implication) can be a consumable resource R denoted with R , e.g. `accepted(ask(Q))`, which holds if R exists in the IS. When a rule fires, R is consumed, i.e. removed from the IS. (If R matches multiple IS terms, only the first matching term is removed.) As an illustration, we can consider the update rule in listing 1 below, stating that if an ask move has been accepted for further processing, the act of responding to the user’s question should be planned (by concluding `agenda(respond(Q, user))`). Note also that `accepted(ask(Q))` is consumed as an effect of firing the rule.

```
integrate_question ::
  accepted(ask(Q)) -*
  agenda(respond(Q, user)).
```

(1)

Antecedents of the form $*R$ always hold; when a rule fires, all resources matching R are consumed.

An antecedent $\wedge R$ expresses that the resource R exists but is not consumed, similar to the *of course* exponential in linear logic (Girard 1987).

An antecedent $\$C$ denotes a condition (Prolog goal), which holds if a proof can be found for C . Importantly, variables can be bound across conditions and linear resources, enabling updates based on linear implication to be integrated

with relations expressed in Prolog. A condition can access resources using the goal `@T`, where T is matched against the content of the IS. Conditions do not modify the IS.

Both antecedents and consequents of update rules can be conjunctions, denoted with Prolog list notation as `[X1, X2, ..., Xn]`. For a more exhaustive specification of the formal rule language, see appendix C.1.

Basic propositions are represented as compound terms, e.g. `satisfied(pat_1)` to signify that patient 1 will be satisfied with surgery. A conjunction of propositions is represented as `[P1, P2, ..., Pn]` where each P_i is a proposition. For example, `[rel_value(back_pain(pat_1), low), rel_value(disability(pat_1), low)]` represents that patient 1 has relatively mild back pain and relatively low disability.

Questions are represented as lambda expressions using SWI Prolog syntax, i.e. `[X1, X2... , Xn]>>Body`. For polar questions, the list of variables is empty. For example, `ask([]>>surgery_recommended(pat_1))` represents the content of the question “Is surgery recommended for the patient?” For wh-questions (including why-questions), the list of variables is non-empty. For example, `[X]>>value(back_pain(pat_1), X)` represents the content of the question “How much back pain does the patient have?”

IS terms can contain variables, denoted with an identifier starting with an uppercase letter or underscore. For example, the term `supports(rel_value(back_pain(X), low), rel_prob(satisfied(X), high), association)` represents that if an individual X has mild back pain, then this supports, by association, that X has a high probability of being satisfied with surgery.

6.2.1 Representing information about the predictive model

Claims corresponding to probabilistic classifications are assumed to be extracted to a term of the form `rel_prob(E,R)` where E denotes an event and R is high if $\mathbb{E}[Y | X] > \mathbb{E}[Y | R]$, otherwise low. For example, if the outcome Y concerns satisfaction with surgery, X describes patient 1, and $\mathbb{E}[Y | X] > \mathbb{E}[Y | R]$, then the term `rel_prob(satisfied(pat_1), high)` is extracted.³ Note that the reference point R (e.g. corresponding to a mean patient) is assumed to be grounded between interlocutors and is not explicitly modelled in terms representing arguments (as will be further discussed in section 6.5.4).

³Neutral claims such as the one in excerpt 4.13 can be accounted for by trichotomising the probability range, as will be discussed in section 6.3.6. Also note that this kind of relative dichotomisation differs from the frequently used maximum a posteriori rule. For example, the popular Python library Scikit-learn (Pedregosa et al. 2011) uses maximum a posteriori in the implementation of the predict method of linear classifiers, which, for binary classification, yields a fixed probability threshold of 0.5.

Data and warrants are assumed to be extracted from the predictive model using the method proposed in chapter 5. Terms representing feature values are extracted and represented as `value(P, V)`, e.g. `value(back_pain(pat_1), 4)`. Terms for relative numerical feature value are extracted using eq. (5.IV) into terms of the form `relative_value(P, R)`, e.g. `relative_value(back_pain(pat_1), low)`.

Warrants are extracted as follows. Given a numerical feature F (coefficient index i) and an outcome Y , the following terms are obtained:

- If $\beta_i > 0$:
`supports(rel_value(F(X), high), rel_prob(Y(X), high), association)` and
`supports(rel_value(F(X), low), rel_prob(Y(X), low), association)`
- If $\beta_i < 0$:
`supports(rel_value(F(X), low), rel_prob(Y(X), high), association)` and
`supports(rel_value(F(X), high), rel_prob(Y(X), low), association)`

For example, if Y is `satisfied(X)`, F is `back_pain` and $\beta_i < 0$, the following terms are extracted:

- `supports(rel_value(back_pain(X), low), rel_prob(satisfied(X), high), association)`
- `supports(rel_value(back_pain(X), high), rel_prob(satisfied(X), low), association)`

For binary features, terms are extracted for both values. Specifically, given a binary feature F (coefficient index i) and an outcome Y , the following terms are obtained:

- If $\beta_i > 0$:
`supports(value(F(X), true), rel_prob(Y(X), high), association)`
and `supports(value(F(X), false), rel_prob(Y(X), low), association)`
- If $\beta_i < 0$:
`supports(value(F(X), true), rel_prob(Y(X), low), association)`
and `supports(value(F(X), false), rel_prob(Y(X), high), association)`

For example, if Y is `satisfied(X)`, F is `has_other_illness` and $\beta_i < 0$ then the following term are extracted:

- `supports(value(has_other_illness(X), true), rel_prob(satisfied(X), low), association)`
- `supports(value(has_other_illness(X), false), rel_prob(satisfied(X), high), association)`

As for categorical features, given a feature F with value V (coefficient index i) and an outcome Y , the following terms are obtained:

- If $\beta_i > 0$:
`supports(value(F(X), V), rel_prob(Y(X), high), association)`
- If $\beta_i < 0$:
`supports(value(F(X), V), rel_prob(Y(X), low), association)`

For example, if Y is `satisfied(X)`, F is `clinic_type`, V is `public`, and $\beta_i < 0$ then the following term is extracted: `supports(value(clinic_type(X), public), rel_prob(satisfied(X), low), association)`.

Warrant terms are ordered in the IS so as to reflect the magnitude of feature contributions. This way, the most influential data are identified first during search and thereby prioritised when conveying multiple explanantia (see section 6.3.3).

It should be noted that predicates representing information about feature values, model predictions and associations learned by the predictive model are intended to reflect propositional attitudes of the explainer. For example, the proposition `rel_prob(satisfied(pat_1), high)` represents that the *explainer believes that* patient 1 has a relatively high probability of being satisfied with surgery, even if the attitudinal aspect is not explicitly encoded in the formal expression. As will be discussed in sections 6.5.6 and 6.5.7, the modelling and management of propositional attitudes has implications in relation to trustworthiness and transparency.

A summary of IS terms containing information about the predictive model is presented in table 6.1.

6.2.2 Representing the state of the dialogue

The interpreted meaning of an explainee's recent utterance U is represented as an IS term having the form `recognized(move(M))`, where M is a dialogue move. A move M to be performed by the system is represented as an IS term of the form `utter(M)`. Once the rule application algorithm has finished, the content of

Table 6.1: Descriptions of IS terms for feature values, model predictions and associations learned by the predictive model.

IS term	Description
$\text{value}(P,V)$	The value of property P is V . A property represents the semantic object to which a value is ascribed and is expressed as an atomic formula, e.g. $\text{back_pain}(\text{pat_1})$ to represent the property “the patient’s back pain”.
$\text{rel_value}(P,R)$	The relative value of property P is R (low, moderate or high).
$\text{prob}(E,P)$	The probability of event E is P .
$\text{rel_prob}(E,R)$	The relative probability of event E is R (low, moderate or high).
$\text{supports}(A,C,M)$	Antecedent A supports consequent C by means M , where M can be e.g. association (A and C are statistically associated) or norm (A supports C by virtue of a norm or convention).

$\text{utter}(M)$ represents the system’s selected next move. The various kinds of moves handled by the dialogue model are summarised in table 6.2 and will be further elaborated in section 6.3.

In addition to terms for input, output and information about the predictive model, the IS also stores information concerning dialogue context. While the specific roles of these terms will be further elaborated in section 6.3 below, they will be briefly introduced here.

The term $\text{accepted}(M)$ represents that the move M has been accepted by the system. This enables the system to pursue different courses of action depending on whether a move is rejected (e.g. if it cannot be answered by the system) or accepted for further processing.

$\text{agenda}(I)$ represents that the agenda item I is planned by the system. An item can be either $\text{respond}(Q, S)$ (respond to question Q raised by speaker S , where S is either user or system), $\text{resume}(J)$ (resume execution of item J), or $\text{argue}(P)$ (provide evidence for proposition P).

$\text{qud}(Q)$ represents that the question Q is currently under discussion (Larsson 2002; Ginzburg 2012). $\text{responded}(Q, P)$ represents that the system has responded to question Q with the proposition P . These pieces of information enable the system to deliver answers incrementally across dialogue turns.

Table 6.2: Descriptions of the kinds of dialogue moves handled by the proposed dialogue model. Modelling of feedback largely follows the taxonomy of interactive communication management (ICM) outlined by Larsson (2002).

Move	Description
ask(Q)	Ask move concerning the question Q .
assert(P)	Assertion of the proposition P .
confirm(P)	Confirmation of the proposition P .
disconfirm(P)	Disconfirmation which asserts the proposition P .
infer(A, C)	Inferential assertion from antecedent A to consequent C , e.g. realised as “since A , C ”.
signal_resumption	Signalling of resumption, e.g. realised as “Also, ...”.
icm(acceptance, positive)	Positive acceptance feedback, e.g. realised as “OK”.
icm(acceptance, negative, R)	Negative acceptance feedback with R as stated reason, where R is either <code>lack_knowledge</code> or a proposition.
icm(understanding, negative, unresolvable_phrase(P))	Negative understand feedback concerning the unresolvable phrase P , e.g. realised as “I don’t understand what ‘heart disease’ means in this context”.
[M1, M2, ..., Mn]	Sequence of moves, where each M_i is a move.

6.3 Dialogue modelling

In the following sub-sections, the proposed model of explanatory dialogue management will be outlined, starting with basic non-explanatory dialogue phenomena which are then incrementally extended with various explanation-related capabilities.

While the dialogue model is intended to be domain-independent, illustrative dialogue examples focus on the domain of AI-assisted treatment choice for spinal stenosis (see chapter 5), where the explainer is a medical specialist (either a doctor or a dialogue system) and the explainee is a human stakeholder, e.g. the patient or a doctor. The dialogue examples assume a scenario involving a specific hypothetical patient (formally identified as `pat_1`) described by the terms in listing 32, and a predictive model described by the terms in listing 33 (see appendix B).

In examples of modelled behaviours, `s` denotes system/explainer and `u` user/

explainee. Contributions are presented both formally (monospaced) and with hypothetical surface forms (italicised). For *u*, formal representations correspond to the sub-term *M* of the IS term `recognized(move(M))`, while for *s*, they correspond to the sub-term *M* of the IS term `utter(M)`.

Some dialogue examples depend on scenario-specific assumptions, presented above a horizontal line. Such terms are assumed to exist in the IS when the dialogue starts.

The entire implementation, including automated tests for all dialogue examples in this chapter (see section 6.4), is available in the accompanying public repository (Berman 2025a).

6.3.1 Basic question answering and assertions

As observed in excerpts 3.2 and 4.15, explainees occasionally ask wh-questions which the explainer can answer by **asserting information which straightforwardly addresses the raised question**. For example, when the respondent in excerpt 4.15 asks for the extraversion score, the operator responds by providing the score. This can be characterised as a basic form of wh-question answering, where the answer is obtained by β -reducing the question.

The examples below illustrate basic wh-question answering for predicted outcome (6.1) and feature value (6.2):

- (6.1) U: `ask([X]»rel_prob(satisfied(pat_1), X))`
Is there a high or low probability that the patient will be satisfied with surgery?
- S: `assert(rel_prob(satisfied(pat_1), high))`
There is a high probability that the patient will be satisfied with surgery.
- (6.2) U: `ask([X]»value(back_pain(pat_1), X))`
How much back pain does the patient have?
- S: `assert(value(back_pain(pat_1), 4))`
The patient's back pain level is 4.

The behaviours in examples 6.1 and 6.2 can be modelled with three update rules. First, a rule marks the interpreted move as accepted for further processing and clears the agenda:

```

mark_move_as_accepted :: [
  recognized(move(Move)),
  *agenda(_)
] -* accepted(Move).

```

(2)

Clearing the agenda when accepting a move ensures that actions planned in relation to previously accepted move are discarded.

Second, a rule integrates the asked question by placing on the agenda the act of responding to the question (see listing 1). Finally, a rule dictates how to respond to a question on the agenda:

```

respond :: [
  agenda(respond(Q, I)),
  $valid_answer(Q, A),
  $answer_move(Q, I, A, M)
] -* utter(M).

```

(3)

The update rule in listing 3 states that if answering question Q raised by interrogator I is on the agenda, and answer move M can be constructed based on a valid answer to Q , then M should be uttered. Conceptually, answer validity entails adherence to the maxims of *quality* (Grice 1975), i.e. that the answer should be held to be true by the speaker⁴, and *relevance*⁵ (Grice 1975) with respect to the question. The relation `valid_answer` is defined as follows:

```

valid_answer([_]>>P, P) :-
  @P.

```

(4)

Listing 4 states that a wh-question with the body P has P as a valid answer if P can be unified with an IS fact. For example, if the IS contains the extracted claim `rel_prob(satisfied(pat_1), high)` and Q is `[X]>>rel_prob(satisfied(pat_1), X)`, then P will be unified with the extracted claim.

Finally, `answer_move(Q, I, P, M)` expresses that given the question Q raised by interrogator I and the answer P , the answer move M can be constructed, and is initially defined as follows:

```

answer_move(_, _, P, assert(P)).

```

(5)

Listing 5 simply states that a question can be answered by asserting the answer.

In the empirical material, explainers occasionally **hedge their claims**. For example, in excerpt 3.3, when the patient asks a (polar) question of whether his lungs can be replaced, the doctor responds with a hedged disconfirmation (“no, I suppose that, unfortunately, it isn’t (feasible)”). In analogy with statistical predictions, it is here assumed that hedging reflects predictive uncertainty. To

⁴Also cf. Searle (1969)’s sincerity condition.

⁵Cf. Ginzburg (1996)’s **about** relation and Larsson (2002)’s **relevant** relation.

capture this behaviour in the dialogue model, the notion of answer validity can be extended so that a polar question can be answered with a relative probability:

```
valid_answer([>>P, rel_prob(P, R)) :-
    @rel_prob(P, R). (6)
```

Furthermore, construction of answer moves needs to be extended accordingly:

```
answer_move(
    [>>P, user, rel_prob(P, high),
    confirm(rel_prob(P, high)))). (7)
```

```
answer_move(
    [>>P, user, rel_prob(P, low),
    disconfirm(rel_prob(P, low)))).
```

Listing 7 states that when a high probability of P can be provided in response to a polar question concerning P , the response can be constructed as a hedged confirmation, while a low probability of P yields a hedged disconfirmation.

With the extensions above, the following dialogue behaviour can be modelled:

```
(6.3) U: ask([>]satisfied(pat_1))
        Will the patient be satisfied with surgery?
        S: confirm(rel_prob(satisfied(pat_1), high))
        Yes, there is a high probability that the patient will be satisfied with surgery.
```

The frequently observed behaviour of **making assertions unprompted** (excerpts 3.1, 3.2, 3.4 to 3.7, 3.9, 3.11, 4.17, 4.23 and 4.31) can be modelled by assuming that an unprompted assertion answers a question that has been placed on the agenda by the system, e.g. in the process of performing a task-specific plan (Larsson 2002) such as informing the patient about potential lab results during a medical consultation. Hence, an unprompted assertion like the one in example 6.4 below is assumed to be triggered by an agenda item (without modelling the mechanism for putting the act of answering the question on the agenda):

```
(6.4) agenda(respond([X]>>rel_prob(satisfied(patient), X), system))


---


        S: assert(rel_prob(satisfied(pat_1), high))
        There is a high probability that the patient will be satisfied with surgery.
```

Since mechanisms for responding to basic wh-questions are already modelled

by the update rules and logic outlined in section 6.3.1, no additional mechanisms need to be introduced in order to account for the behaviour in example 6.4.

6.3.2 Basic explanations

As observed in excerpts 3.1, 3.2, 3.5, 4.10, 4.11, 4.19 and 4.28, interlocutors frequently **explain judgements with reference to specific grounds (data)**, as illustrated by example 6.5 below:

- (6.5) U: ask([], satisfied(pat_1))
Will the patient be satisfied with surgery?
- S: confirm(rel_prob(satisfied(pat_1), high))
Yes, there is a high probability that the patient will be satisfied with surgery.
- U: ask([E, M] supports(E, rel_prob(satisfied(pat_1), high), M))
Why do you think so?
- S: assert(rel_value(back_pain(pat_1), low))
The patient has relatively mild back pain.

The first thing to note about example 6.5 is that U’s anaphoric reference (“Why do you think so?”) to the claim previously asserted by S (that there is a high probability that the patient will be satisfied with surgery) is assumed to be contextually resolved by the NLU component (see section 6.5.8). The content of the contextually resolved ask move is the question [E, M] supports(E, rel_prob(satisfied(pat_1), high), M) which can be expressed as: “By what means *M* does some evidence *E* support that there is a high probability that the patient will be satisfied with surgery?” Here, the explainer responds by asserting a piece of evidence (datum) supporting the claim, which can be modelled with the following extensions:

```
valid_answer(Q, D) :-
  has_variable_and_body(Q, E, supports(E, C, _)),
  @supports(D, C, _),
  @D.
(8)
```

```
has_variable_and_body(Vars>>Body, Var, Body) :-
  member(Var1, Vars),
  Var1 == Var.
```

The answer validity clause in listing 8 states that a question concerning evidence supporting a consequent *C* can be answered with the datum *D*, if *D* and *C* match a support proposition held to be true, and *D* is held to be true.

It can be noted that the matching against support propositions (`@supports(D, C, _)`) uses unification. Hence, a support proposition can be expressed on a more general level than the evidence and consequent, as in example 6.5 above, where the support proposition concerns any potential individual X rather than the specific patient discussed in the dialogue. Also note that even if multiple data can be identified (one for each warrant), the update rule in listing 3 only focuses on the first solution; this restriction will be removed in section 6.3.3.

The extension in listing 8 also provides the helper predicate `has_variable_1` and `and_body(Q, V, B)` which tests whether a question Q with body B has V as one of its variables.

Occasionally (excerpts 3.9, 3.10 and 4.15), interlocutors **explain judgements with reference to general rules or principles (warrants)**. Based on the available empirical material, it cannot be determined how interlocutors decide whether to explain judgements using data or warrants. In order not to build unsubstantiated assumptions into the model, it is here conceived that when there are several valid answers to a question, an answer selection policy can be used for choosing which of the valid answers to include in the response. Answer selection policies are not built into the dialogue model outlined here but are instead modelled as assumptions, as illustrated by example 6.6 below, where the explainer chooses a warrant-based explanation:

```
(6.6) answer_selection_policy(
    [E, M]>>supports(E, _, M),
    Valid,
    [Selected],
    (member(Selected, Valid), Selected=supports(_,-,_)))
```

```
U: ask([ ]>satisfied(pat_1))
    Will the patient be satisfied with surgery?

S: confirm(rel_prob(satisfied(pat_1), high))
    Yes, there is a high probability that the patient will be satisfied with surgery.

U: ask([E, M]>supports(E, rel_prob(satisfied(pat_1), high), M))
    Why do you think so?

S: assert(
    supports(
        rel_value(back_pain(X), low),
        rel_prob(satisfied(X), high),
        association))
    I associate mild back pain with a high probability of being satisfied with surgery.
```

Answer selection policies are formalised as meta-predicates expressed as `answer_selection_policy(Q, Candidates, Selected, Cond)`, where `Candidates` contains all valid answers, `Selected` contains the answers to select, and `Cond` is a condition (Prolog goal) which instantiates `Selected`. The policy assumed in example 6.6 states that for a question concerning the epistemic basis of a claim, the first valid answer which is a support proposition (i.e. a warrant) is selected.

To model the behaviour in example 6.6, the update rule in listing 3 is extended to handle answer selection:

```
respond :: [
    agenda(respond(Q, I)),
    $findall(A, valid_answer(Q, A), ValidAnswers),
    $select_answers(Q, ValidAnswers, SelectedAnswers),
    $answer_move(Q, I, SelectedAnswers, M)
] -* utter(M). (9)
```

The relation `select_answers` is defined as follows:

```

select_answers(Q, Candidates, Result) :-
  (@answer_selection_policy(Q, Candidates, Selected, Condition) ->
   Condition,
   intersection(Candidates, Selected, Result)
  )
;
  Result = Candidates
).

```

(10)

Listing 10 states that if an applicable answer selection policy exists, it should be invoked when selecting answers⁶; if no applicable policy exists, all valid answers are selected.

Furthermore, answer validity is extended to treat warrants as answers:

```

valid_answer(Q, W) :-
  has_variable_and_body(Q, E, supports(E, C, _)),
  W = supports(WA, _, _),
  @W,
  unifiable(WC, C, _),
  copy_term(WA, WA1),
  @WA1.

```

(11)

Listing 11 states that for questions concerning evidence supporting a consequent C , the warrant W , which expresses that the antecedent W_A supports C , is a valid answer if W and its antecedent W_A are both held to be true.

Finally, the definition of `answer_move` provided in listing 5 is extended to handle multiple answers. Specifically, `answer_move(Q, I, Ps, M)` expresses that given the question Q raised by interrogator I and the answers in P , the answer move M can be constructed, and is defined as follows:⁷

```

answer_move(_, _, Ps, assert(P)) :-
  normalize(Ps, P).

normalize([X], X) :- !.
normalize(X, X).

```

(12)

It should be noted that, as a consequence of these extensions to the model, the frequently observed enthymematic behaviour for datum-based explanantia, illustrated by example 6.5, needs to assume a policy for selecting a datum answer:

⁶By returning the intersection between candidate answers and those selected by the policy, it is ensured that invalid (insincere or irrelevant) answers returned by the policy are excluded.

⁷By convention, Prolog list variables are named with a trailing *s* (e.g. Ps) and denote sequences, which are represented in formal notation by bold symbols (e.g. P).

```

answer_selection_policy(
  [E, M]>>supports(E, _, M),
  Valid,
  [Selected],
  (member(Selected, Valid), Selected\=supports(.,.,.)))

```

(13)

In the examples of basic explanatory behaviours discussed above, explanations are provided on demand. However, in the observed human dialogues, interlocutors occasionally **explain a claim unprompted when delivering it** (excerpts 3.1, 3.6 and 3.7). Such utterances are here modelled as dialogue moves conveying an inference from evidence to conclusion. It is assumed that the agenda can contain an item for arguing for a claim C (without modelling the reasons for how such an item ends up on the agenda), as illustrated by example 6.7:⁸

```

(6.7) agenda(argue(rel_prob(satisfied(pat_1), high)))
      inherit select_1st_datum

```

```

S: infer(
    rel_value(back_pain(pat_1), low),
    rel_prob(satisfied(pat_1), high))

```

Since the patient has relatively mild back pain, there is a high probability that the patient will be satisfied with surgery.

The behaviour in example 6.7 can be modelled by adding an update rule for processing argue items on the agenda:

```

argue :: [
  agenda(argue(C)),
  $(Q = [E, M]>>supports(E, C, M)),
  $findall(A, valid_answer(Q, A), ValidAnswers),
  $select_answers(Q, ValidAnswers, SelectedAnswers),
  $normalize(SelectedAnswers, A)
] -* utter(infer(A, C)).

```

(14)

Listing 14 expresses that if there is an agenda item for providing evidence for the claim C , and A contains evidence supporting C , then an inference from A to C should be uttered.

6.3.3 Complex explanantia

Explanatory behaviours modelled thus far concern situations where the explainer supports a claim with a single datum or warrant. However, in the studied human dialogues (see excerpts 3.1, 3.10, 4.10, 4.17 and 4.19), judgements are frequently

⁸select_1st_datum refers to the policy in listing 13. Generally, `inherit` R imports assumptions from the resource R.

justified with complex explanantia involving multiple pieces of evidence. In the empirical material, complex explanantia are delivered either in a single turn, or incrementally across multiple turns in joint coordination between interlocutors. The former strategy can be modelled by assuming a policy for selecting multiple data, as illustrated for multiple pieces of data in example 6.8 below:

```
(6.8) answer_selection_policy(
      [E, M]>>supports(E, _, M),
      Valid,
      Selected,
      include([A]>>(A \= supports(_, _, _)), Valid, Selected))
```

- U: ask([]>satisfied(pat_1))
Will the patient be satisfied with surgery?
- S: confirm(rel_prob(satisfied(pat_1), high))
Yes, there is a high probability that the patient will be satisfied with surgery.
- U: ask([E, M]>supports(E, rel_prob(satisfied(pat_1), high), M))
Why do you think so?
- S: assert([
 rel_value(back_pain(pat_1), low),
 rel_value(disability(pat_1), low),
 value(has_other_illness(pat_1), false)
])
The patient has relatively mild back pain, low disability and no other illnesses.

The policy in example 6.8 (hereafter referred to with the resource `select_1` data) states that for a question concerning the epistemic basis of a claim, all valid data are selected.

Incremental delivery can be modelled by assuming a policy which constrains the amount of selected answers to a suitable turn-constructional unit (Sacks, Schegloff and Jefferson 1974), e.g. a maximum of one datum, as in example 6.9 below:⁹

⁹Similarly, only selecting the first warrant yields a behaviour similar to the one in excerpt 3.10, where warrants are delivered incrementally.

(6.9) inherit select_1st_datum

U: ask([]»satisfied(pat_1))
Will the patient be satisfied with surgery?

S: confirm(rel_prob(satisfied(pat_1), high))
Yes, there is a high probability that the patient will be satisfied with surgery.

U: ask([E, M]»supports(E, rel_prob(satisfied(pat_1), high), M))
Why do you think so?

S: assert(rel_value(back_pain(pat_1), low))
The patient has relatively mild back pain.

U: icm(acceptance, positive)
I see.

S: signal_resumption
assert(rel_value(disability(pat_1), low))
Also, the patient has relatively low disability.

U: icm(acceptance, positive)
OK.

S: -
(Does not take turn)

To support incremental delivery, a few extensions are required. First, the rule for responding to questions is extended to track previously delivered responses via an additional consequent:

```
respond :: [  
  agenda(respond(Q, I)),  
  $findall(A, valid_answer(Q, A), ValidAnswers),  
  $select_answers(Q, ValidAnswers, SelectedAnswers),  
  $answer_move(Q, I, SelectedAnswers, M)                                (15)  
] -* [  
  utter(Move),  
  responded(Q, SelectedAnswers)  
].
```

Second, questions under discussion need to be tracked. Specifically, when a

question Q is integrated, $\text{qud}(Q)$ is concluded:¹⁰

```
integrate_question :: [
  accepted(ask(Q)),
  *responded(Q, _)
] -* [
  qud(Q),
  agenda(respond(Q, user))
].
```

(16)

Third, when answer delivery is met with positive acceptance, the agent can conclude that the question under discussion should be responded to by resuming previous delivery:

```
integrate_acknowledgement :: [
  accepted(icm(acceptance, positive)),
  ^qud(Q)
] -* agenda(resume(respond(Q, user))).
```

(17)

Finally, an update rule is added for selecting answers when resuming a response:

```
resume_responding :: [
  agenda(resume(respond(Q, I))),
  $findall(A, (
    valid_answer(Q, A),
    \+ (@responded(Q, As), member(A, As))
  ), ValidAnswers),
  $select_answers(Q, ValidAnswers, SelectedAnswers),
  $(SelectedAnswers \== []),
  $answer_move(Q, I, SelectedAnswers, Move)
] -* [
  utter([signal_resumption, Move]),
  responded(Q, SelectedAnswers)
].
```

(18)

Unlike normal response selection, previously delivered answers to the present question are excluded (by utilising Prolog's negation operator $\backslash+$), and the answer move is preceded by a resumption marker.

It can be noted that, by default, the sequence in which explanantia are mentioned by the system is determined by the ordering of support propositions, which in turn reflects local feature importance (see section 6.2.1).

¹⁰The antecedent $\text{*responded}(Q, _)$ ensures that when a question Q is integrated, any potential previous responses to Q are consumed. This way, if a question is explicitly re-raised, the agent re-states its answers.

6.3.4 Chained inferences

As observed in section 3.4.5, a judgement supported by multiple inference steps can be explained by highlighting evidence which is inferentially remote from the conclusion. This behaviour can be modelled with a recursive notion of answer validity, where an answer A to a question concerning the epistemic basis of a claim C is considered valid either if

- A can be linked with the C via a warrant (as stated in listing 8), *or*
- an intermediate piece of evidence E can be linked with C via a warrant, and A explains the epistemic basis of E .

This recursive definition accounts for inference chains of arbitrary length, and can be formalised by modifying the previously outlined clause for answer validity in listing 8 as follows:

```
valid_answer(Q, D) :-
    has_variable_and_body(Q, E, supports(E, C, _)),
    supports_directly_or_indirectly(D, C).

supports_directly_or_indirectly(A, C) :-
    @supports(A, C, _),
    @A. (19)

supports_directly_or_indirectly(A, C) :-
    @supports(A1, C, _),
    @A1,
    supports_directly_or_indirectly(A, A1).
```

As a consequence of the modification in listing 19, all data in an inference chain will be considered valid. However, in the observed excerpt 3.2, where the doctor explains his intention to have the patient's lungs x-rayed with reference to the patient's low oxygen saturation, only the inferentially most distant datum is elicited. To account for this specific behaviour, a policy for preferring empirically grounded evidence is assumed, as illustrated by example 6.10 below:

```
(6.10) supports(
  rel_prob(satisfied(X), high),
  surgery_recommended(X),
  norm)
answer_selection_policy(
  [E, M]>>supports(E, _, M),
  Valid,
  Selected,
  (member(D, Valid),
   (D=rel_value(_, _) ; D=value(_,_)) ->
   Selected = [D]
  ; Selected = Valid))
```

U: ask([E, M]>>supports(E, surgery_recommended(pat_1), M))
Why do you recommend surgery for the patient?

S: assert(rel_value(back_pain(pat_1), low))
The patient has relatively mild back pain.

On an abstract level, the argumentative structure of example 6.10 is identical to that of excerpt 3.2: A claim is explained with reference to a datum which supports an intermediate conclusion which, in turn, supports the claim whose epistemic basis is interrogated. The agent's selection of the inferentially remote datum is here assumed to be caused by a preference for highlighting empirically grounded facts over probabilistic judgements. This is formalised by the selection policy in example 6.10, which states that if a datum (relative or absolute value) exists among the valid answers, that datum is selected; otherwise, all valid answers are selected. This strategy can be contrasted with the variant in example 6.11 below (corresponding to excerpt 3.2''), where the agent instead focuses on the inferentially most proximate datum:

```
(6.11) supports(
    rel_prob(satisfied(X), high),
    surgery_recommended(X),
    norm)
inherit select_1st_datum
```

U: ask([E, M]»supports(E, surgery_recommended(pat_1), M))
Why do you recommend surgery for the patient?

S: assert(rel_prob(satisfied(pat_1), high))
There is a high probability that the patient will be satisfied with surgery.

U: ask([E, M]»supports(E, rel_prob(satisfied(pat_1), high), M))
Why do you think so?

S: assert(rel_value(back_pain(pat_1), low))
The patient has relatively mild back pain.

6.3.5 Forward expansions

In excerpt 3.5, when a doctor presents evidence (a datum) and the patient responds by asking what the evidence means, the doctor delivers a conclusion based on the bespoke evidence. Argumentatively, the patient's question can be characterised as a request for forward expansion (Xydis et al. 2020). To model such behaviour, one first of all needs to account for the explainer's reason for asserting the evidence. As in section 6.3.1, it is here assumed that the assertion is triggered by an item which already has been placed on the agenda. Example 6.12 below illustrates the desired behaviour:

```
(6.12) agenda(respond([X]>>rel_value(back_pain(patient), X), system))
inherit select_1st_datum
```

S: assert(rel_value(back_pain(pat_1), low))
The patient has relatively mild back pain.

U: ask([C, M]»supports(rel_value(back_pain(pat_1), low), C, M))
What might this imply?

S: assert(rel_prob(satisfied(pat_1), high))
I think that the patient has a high probability of being satisfied with surgery.

To account for the behaviour in example 6.12, the notion of answer validity is extended to cover questions concerning which conclusion that can be drawn from evidence, as formalised in listing 20:

```
valid_answer(Q, C) :-
  has_variable_and_body(Q, XC, supports(D, XC, _)),
  supports_directly_or_indirectly(D, C). (20)
```

Listing 20 states that C is a valid answer to a question concerning which conclusion that can be drawn from the datum D , if D (directly or indirectly) supports C .

6.3.6 Contradictory and neutral evidence

While most claims in the observed dialogues are explained with evidence that one-sidedly supports the conclusion, two exceptions can be observed: explanations for numerical and neutral claims. The former is observed in excerpt 4.9, where the operator’s numerical framing of the prediction (“You scored 5 on Openness”) is also picked up in the respondent’s explanation question (“Why is it a 5 on Openness?”). As reflected in the response from the operator and illustrated by example 6.13 below, numerical claims can be explained with contradictory evidence:

```
(6.13) U: ask([X]»prob(satisfied(pat_1), X))
         How probable is it that the patient will be satisfied with surgery?

S:      assert(prob(satisfied(pat_1), 0.75))
         There is a 75% probability that the patient will be satisfied with surgery.

U:      ask([E, M]»supports(E, prob(satisfied(pat_1), 0.75), M))
         Why do you think so?

S:      infer(
         rel_value(disability(pat_1), low),
         rel_prob(satisfied(pat_1), high)),
         infer(
         rel_value(back_pain(pat_1), high),
         rel_prob(satisfied(pat_1), low))
         The patient’s relatively mild back pain, low disability and absence of other illnesses indicate that the patient will be satisfied with surgery, while the patient’s relatively long duration of back and leg pain indicates that the patient will be unsatisfied.
```

To model the behaviour in example 6.13, the notion of answer validity can be extended so that a question concerning the epistemic basis of a probability of an event E can be answered with evidence that explains why the relative probability of E is either high or low:

```
valid_answer(Vars>>supports(A, prob(E, _), M), Ans) :-
    member(R, [low, high]),
    valid_answer(Vars>>supports(A, rel_prob(E, R), M), Ans).      (21)
```

Furthermore, in order to indicate whether evidence points in positive or negative direction (rather than just stating the evidence as such), answer moves for conflicting evidence can be constructed as inferences from evidence to conclusion:

```
answer_move(Vars>>supports(E, C, M), _, As, Move) :-
    member(C, [prob(P, _), rel_prob(P, moderate)]),
    findall(
        infer(NormalizedEvidence, rel_prob(P, R)),
        (
            member(R, [high, low]),
            findall(A,
                (member(A, As),
                 valid_answer(Vars>>supports(E, rel_prob(P, R), M), A)),
                Evidence),
            Evidence \== [],
            normalize(Evidence, NormalizedEvidence)
        ),
        Inferences),
    Inferences \== [],
    normalize(Inferences, Move).
```

Listing 22 states that a question concerning the epistemic basis of a probability (or a moderate/neutral relative probability, see below) can be answered using inferences from evidence to a conclusion that the relative probability is high or low.

Neutral claims are observed in excerpts 4.12 (“on extroversion, you scored *pretty close to the median*”) and 4.13 (“(you have) *neutral* on the rest”). In the latter case, an explanation is conveyed for the prediction, namely by referring to neutral feature values (“there is a strong relationship between extroversion and dancability and liveliness”, “your preferences didn’t indicate strongly one way or another about those features”). Neutrally framed claims can be modelled with an intermediate (moderate) range for relative feature values and probabilities, as exemplified below:

```
(6.14) rel_value(back_pain(pat_1), moderate)
rel_value(disability(pat_1), moderate)
rel_prob(satisfied(pat_1), moderate)
agenda(respond([]>>satisfied(pat_1), system))
inherit select_data
```

```
S:   assert(rel_prob(satisfied(pat_1), moderate))
      The patient has a moderate probability of being satisfied with
      surgery.

U:   ask([[E, M]]»supports(E, rel_prob(satisfied(pat_1), moderate), M))
      Why do you think so?

S:   assert([
      rel_value(back_pain(pat_1), moderate),
      rel_value(disability(pat_1), moderate)])
      The patient has moderate back pain and disability.
```

The behaviour in example 6.14 can be modelled by encoding that moderate feature values support a moderate estimate of outcome;¹¹ specifically, the argument extraction method outlined in section 6.2.1 can be extended so that for each numerical feature F (coefficient index i) and outcome Y , the term `supports(rel_value(F(X), moderate), rel_prob(Y(X), moderate), association)` is extracted.

A neutral claim can also be supported by contradictory evidence, as exemplified by excerpt 4.14 and illustrated by the corresponding example below:

¹¹Mathematically, this assumes that $X \approx R \implies \mathbb{E}[Y | X] \approx \mathbb{E}[Y | R]$, which is typically expected for GLMs given the smoothness of $\mathbb{E}[Y | X]$ w.r.t. X .

```
(6.15) rel_value(disability(pat_1), low)
rel_value(back_pain(pat_1), high)
rel_prob(satisfied(pat_1), moderate)
agenda(respond([]>>satisfied(pat_1), system))
inherit select_data
```

```
S: assert(rel_prob(satisfied(pat_1), moderate))
    The patient has a moderate probability of being satisfied with surgery.

U: ask([E, M]»supports(E, rel_prob(satisfied(pat_1), moderate), M))
    Why do you think so?

S: infer(
    rel_value(disability(pat_1), low),
    rel_prob(satisfied(pat_1), high)),
infer(
    rel_value(back_pain(pat_1), high),
    rel_prob(satisfied(pat_1), low))
    The patient's low disability indicates that the patient will be satisfied with surgery, while the patient's relatively long duration of back indicates that the patient will be unsatisfied.
```

To support the behaviour in example 6.15, the notion of answer validity can be extended to handle conflicting evidence for moderately probable events:

```
supports_directly_or_indirectly(A, C) :-
    C = rel_prob(Event, moderate),
    @C,
    findall(
        E,
        (
            member(R, [high, low]),
            supports_directly_or_indirectly(E, rel_prob(Event, R))
        ),
        Evidence),
    member(A, Evidence).
(23)
```

Listing 23 states that a moderate assessment concerning the probability of event E can be supported by both positive and negative evidence for E .

It should be noted that while polar distinctions between low and high can be evaluated objectively in relation to a reference, a third, neutral value for relative feature values and probabilities introduces additional elements of judgement which somehow need to be accounted for. In multivariate situations (such as in examples 6.14 and 6.15), this becomes non-trivial due to dependencies between judgements. For example, if feature A is below and feature B above the reference level, but neither by much, then in order to explain a moderate estimate of

outcome, A can be characterised as low and B as high, or both as moderate; however, judging one of them as low/high and the other as moderate does not seem to intuitively explain the neutral conclusion, even if both judgements seem reasonable in isolation. The approach outlined above, which delegates mutually dependent judgements to the argument extraction method without modelling how such judgements are performed, leaves these matters unresolved.

6.3.7 Framing of data

The empirical analysis revealed three ways in which numerical data is conveyed in support of claims: relatively (as discussed in section 6.3.2 and illustrated by several examples above), numerically (as in excerpt 3.8), or through a combination thereof (as in excerpt 3.1). To model all of these cases, the second clause in listing 19 can be modified as follows:¹²

```
supports_directly_or_indirectly(D, C) :-
  @supports(A, C, _),
  @A,
  (
    D = A
    ;
    A = rel_value(P, _),
    D = value(P, _),
    @D
  ).
```

(24)

Listing 24 states that a question concerning evidence supporting a consequent C can be answered *either* with an antecedent A supporting C *or* an exact value which elaborates a relative value supporting C.

Examples 6.16 to 6.18 below illustrate how the different kinds of observed strategies for conveying data can be modelled using different policies (with differences in surface forms emphasised).

¹²As a consequence, the definition of the resource `select_data` needs to be changed to the one in example 6.16. A similar change needs to be done for `select_1st_datum`.

```
(6.16) answer_selection_policy(
  [E, M]>>supports(E, _, M),
  Valid,
  Selected,
  findall(A, (
    member(A, Valid),
    (A=rel_value(_,_);
     A=value(P,_), \+ member(rel_value(P,_), Valid))
  ), Selected)
)
```

U: ask([E, M]»supports(E, rel_prob(satisfied(pat_1), high), M))
Why do you think that the patient has a high probability of being satisfied with surgery?

S: assert([
 rel_value(back_pain(pat_1), low),
 rel_value(disability(pat_1), low),
 value(has_other_illness(pat_1), false)
])
The patient has relatively mild back pain, relatively low disability, and no other illnesses.

```
(6.17) answer_selection_policy(
  [E, M]>>supports(E, _, M),
  Valid,
  Selected,
  include([A]>>(A = value(_, _)), Valid, Selected))
```

U: ask([E, M]»supports(E, rel_prob(satisfied(pat_1), high), M))
Why do you think that the patient has a high probability of being satisfied with surgery?

S: assert([
 value(back_pain(pat_1), 4),
 value(disability(pat_1), 26),
 value(has_other_illness(pat_1), false)
])
The patient has a back pain of 4, a disability of 26, and no other illnesses.

```
(6.18) answer_selection_policy(
  [E, M]>>supports(E, _, M),
  Valid,
  Selected,
  include([A]>>(A \= supports(_, _, _)), Valid, Selected))
```

U: ask([E, M]>>supports(E, rel_prob(satisfied(pat_1), high), M))
Why do you think that the patient has a high probability of being satisfied with surgery?

S: assert([
 rel_value(back_pain(pat_1), low),
 value(back_pain(pat_1), 4),
 rel_value(disability(pat_1), low),
 value(disability(pat_1), 26),
 value(has_other_illness(pat_1), false)
])

The patient's back pain is only 4 and disability only 26, and the patient has no other illnesses.

6.3.8 Confirmation questions

Two kinds of polar/confirmation questions concerning the relation between data and claim have been encountered: whether hypothetical data support a previously asserted claim or its counter-claim (excerpt 4.18 and rewritten excerpt 3.12'n) and whether previously asserted data support a hypothetical claim (excerpts 3.4 and 3.5). Modelling of the former is illustrated by example 6.19 below:

(6.19) agenda(respond([]>>satisfied(pat_1), system))

S: assert(rel_prob(satisfied(pat_1), high))
There is a high probability that the patient will be satisfied with surgery.

U: ask([]>>
 supports(
 rel_value(back_pain(pat_1), low),
 rel_prob(satisfied(pat_1), high),
 _))
Is this because the patient has mild back pain?

S: confirm(
 supports(
 rel_value(back_pain(pat_1), low),
 rel_prob(satisfied(pat_1), high),
 association))
Yes, the patient's relatively mild back pain supports the assessment that the patient will be satisfied with surgery.

The latter kind of question is illustrated by example 6.20 below:

```
(6.20) agenda(
  respond(
    [E, M]>>supports(
      E, rel_prob(satisfied(pat_1), high), M), system))
inherit select_data
```

```
S:  assert([
      rel_value(back_pain(pat_1), low),
      rel_value(disability(pat_1), low),
      value(has_other_illness(pat_1), false)
    ])
    The patient has relatively mild back pain, low disability and no
    other illnesses.
```

```
U:  ask([]>>
      supports(
        [
          rel_value(back_pain(pat_1), low),
          rel_value(disability(pat_1), low),
          value(has_other_illness(pat_1), false)
        ],
        rel_prob(satisfied(pat_1), high),
        _))
    Does this suggest that the patient will be satisfied with surgery?
```

```
S:  confirm(
      supports(
        [
          rel_value(back_pain(pat_1), low),
          rel_value(disability(pat_1), low),
          value(has_other_illness(pat_1), false)
        ],
        rel_prob(satisfied(pat_1), high),
        _))
    Yes, the patient's relatively mild back pain, low disability and
    absence of other illnesses support the assessment that the patient
    will be satisfied with surgery.
```

A somewhat more intricate situation emerges when only a subset of the data highlighted by the explainee actually supports the conclusion, as in example 6.21 below:

```
(6.21) agenda(
  respond(
    [E, M]>>supports(
      E, rel_prob(satisfied(pat_1), high), M), system))
```

S: assert(rel_prob(satisfied(pat_1), high))
There is a high probability that the patient will be satisfied with surgery.

U: ask([]>>
 supports(
 [
 rel_value(back_pain(pat_1), low),
 rel_value(back_pain_duration(pat_1), high)
],
 rel_prob(satisfied(pat_1), high),
 _))
Is this because the patient has had mild back pain for a long duration of time?

S: disconfirm(
 not(
 supports(
 [rel_value(back_pain_duration(pat_1), high)],
 rel_prob(satisfied(pat_1), high), _)))
No, the patient's long duration of back pain does not support the assessment that the patient will be satisfied with surgery.

An explainer's ability to answer these kinds of polar questions is not necessarily restricted to claims which are held to be true by the explainer. Example 6.22 below illustrates the ability of an explainer to confirm counter-evidence:

```
(6.22) agenda(
  respond(
    [E, M]>>supports(
      E, rel_prob(satisfied(pat_1), high), M), system))
```

S: confirm(rel_prob(satisfied(pat_1), high))
There is a high probability that the patient will be satisfied with surgery.

U: ask([]>>
 supports(
 rel_value(back_pain_duration(pat_1), high),
 rel_prob(satisfied(pat_1), low),
 _))
Does the patient's long duration of back pain point in the other direction?

S: confirm(
 supports(
 rel_value(back_pain_duration(pat_1), high),
 rel_prob(satisfied(pat_1), low),
 association))
Yes, the patient's relatively long duration of back pain supports the assessment that the patient will be unsatisfied with surgery.

An additional type of polar question can potentially be observed in excerpt 3.2, where the patient's question concerning whether low oxygen uptake can indicate poor lung function can be characterised as a request for confirming a hypothetical warrant (as captured by the rewritten excerpt 3.2'). This behaviour is illustrated by example 6.23 below:

```
(6.23) U: ask([]>>
          supports(
            rel_value(back_pain(X), high),
            rel_prob(satisfied(X), high),
            -))
```

Does severe back pain indicate that a patient will be satisfied with surgery?

```
S: disconfirm(
    not(
      supports(
        rel_value(back_pain(X), high),
        rel_prob(satisfied(X), high),
        association)))
```

No, I don't associate severe back pain with a high probability of being satisfied with surgery.

To account for the behaviours above, the notion of answer validity can be extended as follows:

```
valid_answer([]>>P, P) :-
  @P.

valid_answer([]>>P, not(P)) :-
  @not(P).

valid_answer([]>>supports(Es, C, M), supports(Es, C, M)) :-
  is_list(Es),
  forall(
    member(E, Es),
    supports_directly_or_indirectly(E, C)).
```

(25)

```
valid_answer([]>>supports(Es, C, M), not(supports(Zs, C, M))) :-
  findall(
    Z,
    (member(Z, Es), \+ supports_directly_or_indirectly(Z, C)),
    Zs),
  Zs \== [].
```

The first two clauses in listing 25 enable straightforward confirmations and disconfirmations for propositions that the system holds as true and false respectively. The other two clauses deal with questions concerning whether a conjunction of evidence E supports a conclusion C . Specifically, the third clause states that such a question can be answered positively if each piece of evidence $E \in E$ supports C , while the fourth clause states that it can be answered with the negative counter-examples Z if some pieces of evidence in fact do not support the conclusion.

Additionally, construction of answer moves needs to be extended to handle straightforward confirmations and disconfirmations (first two clauses below) as well as negative counter-examples for complex explanantia (third clause below):

```
answer_move([>>P, user, [P], confirm(P)).

answer_move([>>P, user, [not(P)], disconfirm(not(P)))).

answer_move(
  [>>supports(Es, C, M),
  user,
  [not(supports(Zs, C, M))],
  disconfirm(not(supports(Zs, C, M)))).
```

(26)

Furthermore, in addition to extracting positive warrants from the statistical model (as described in section 6.2.1), negative warrants also need to be extracted. For example, if $\beta_i < 0$ for back pain, the following terms are extracted:

- not(supports(rel_value(back_pain(X), high), rel_ prob(satisfied(X), high), association)) (severe back pain is not statistically associated with a high probability of being satisfied with surgery)
- not(supports(rel_value(back_pain(X), low), rel_ prob(satisfied(X), low), association)) (mild back pain is not statistically associated with a low probability of being satisfied with surgery)

6.3.9 Presupposition violations

The empirical analysis revealed two instances (excerpts 4.26 and 4.27) where the explainer signals a presupposition violation associated with the explainee's request for an explanation. Modelling of such behaviour can be decomposed into two steps: identification of presuppositions in a given utterance, and signalling of potential presupposition violations. Identification of presuppositions is here conceived as a problem related primarily to natural language understanding, as illustrated by examples 6.24 and 6.25 below, where presuppositions triggered by U's utterances are represented as a separate IS term recognized(presupposition(P)):

(6.24) `recognized(presupposition(rel_prob(satisfied(pat_1), low)))`

U: `ask([E, M]»supports(E, rel_prob(satisfied(pat_1), low), M))`
Why do you think that the patient has a low probability of being satisfied with surgery?

S: `icm(acceptance, negative, not(rel_prob(satisfied(pat_1), low)))`
The patient's probability of being satisfied is not low.

(6.25) `recognized(presupposition(rel_value(back_pain(pat_1), low)))`

U: `ask([]>> supports(rel_value(back_pain(pat_1), high), rel_prob(satisfied(pat_1), high), _))`
Does the patient's severe back pain indicate that the patient will be unsatisfied with surgery?

S: `icm(acceptance, negative, not(rel_value(back_pain(pat_1), high)))`
The patient does not have severe back pain.

To model signalling of presupposition violations, an update rule is added which rejects a move if it contains a false presupposition:

```
reject_move_with_presupposition_violation :: [  
  recognized(presupposition(Presupposition)),  
  ^Belief,  
  $contradicts(Belief, Presupposition),  
  recognized(move(_))  
] -* utter(icm(acceptance, negative, not(Presupposition))).
```

(27)

The rule in listing 27 states that if a move triggers a presupposition which is contradicted by any of the agent's beliefs, negative acceptance should be signalled. The notion of contradiction is modelled as follows:

```
contradicts(rel_prob(E, X), rel_prob(E, Y)) :-  
  X \== Y.  
  
contradicts(rel_value(P, X), rel_value(P, Y)) :-  
  X \== Y.
```

(28)

The first clause in listing 28 establishes that a proposition stating that the

relative probability of event E is X contradicts a proposition stating that the relative probability of E is something else than X . The second clause establishes the corresponding relation for relative values.

It should be noted that questions embedded in U's utterances in examples 6.24 and 6.25 can, in principle, be raised without triggering presupposition violations (and answered using the strategies modelled thus far). For example, the ask move conveyed by an utterance such as "Are there any indications that the patient would be unsatisfied with surgery?" can be formalised identically as the one in example 6.24, without presupposing that S believes that the patient will be unsatisfied with surgery.

6.3.10 Answer unavailability

The empirical analysis revealed various types of scenarios where it might be useful for a CXAI system to signal answer unavailability. In one such scenario (see the rewritten excerpt 4.20'), the explainee asks a question concerning a type of evidence which the explainer understands semantically but whose potential relation to the explanandum is unknown to the explainer, as illustrated by example 6.26 below:

(6.26) U: ask([X]>>
 supports(
 has_other_specific_illness(pat_1, X),
 rel_prob(satisfied(pat_1), low),
 _))

Which of the patient's other illnesses indicate that the patient will be unsatisfied with surgery?

S: icm(acceptance, negative, lack_knowledge)
 I don't have any information about that.

Example 6.26 assumes that S is able to semantically represent the concept `has_other_specific_illness(P, X)` (P has another specific illness X) but lacks knowledge concerning how such propositions relate to satisfaction with surgery.

Another observed scenario (excerpt 4.5) concerns questions about causality, as illustrated by example 6.27 below:

```
(6.27) U: ask([]>>
          supports(
            rel_value(back_pain(X), _),
            rel_prob(satisfied(X), high),
            causation))
          Is amount of back pain causally related to satisfaction with surgery?

S: icm(acceptance, negative, lack_knowledge)
   I don't have any information about that.
```

Both kinds of scenarios can be modelled by adding an update rule for rejecting unanswerable questions:

```
reject_unanswerable_question :: [
  recognized(move(ask(Q))),
  $(\+ valid_answer(Q, _))
] -* utter(icm(acceptance, negative, lack_knowledge)). (29)
```

The update rule in listing 29 states that if an ask move for the question Q has been recognised, but no valid answers are available for Q , negative acceptance should be signalled, with lack of knowledge as the reason.

It can be noted that this approach to modelling the ability to signal answer unavailability does not require that the agent has information regarding the reasons for *why* it lacks knowledge. For example, the behaviour illustrated in example 6.27 does not require that the agent knows what the means causation is, or that information about causality cannot be derived from a statistical model. Instead, the modelled mechanism draws merely on an ability to detect that a valid answer cannot be identified given the agent's available knowledge.

6.3.11 Negative understanding

Generally, an explainer's ability to respond to an explanation request requires the explainer to map concepts in the explainee's utterance onto the explainer's internal knowledge representations. In the examples discussed above, this condition has been assumed to be fulfilled. However, for some of the analysed human dialogues, it might be unrealistic to assume that a CXAI system would be able to perform such a mapping (see the rewritten excerpts 4.11' and 4.12'). In such scenarios, it might be adequate for the explainer to signal negative understanding, as illustrated by example 6.28 below:

(6.28) `recognized(contextually_unresolvable_phrase("heart disease"))`

U: *Does the patient's heart disease indicate that the patient will be unsatisfied with surgery?*

S: `icm(understanding, negative, unresolvable_phrase("heart disease"))`

Sorry, I don't understand what "heart disease" means in this context.

In example 6.28, U asks a question involving the phrase "heart disease", which S might understand lexicosemantically, but not in relation to its internal representations about the domain. This lack of understanding is captured by the IS term `recognized(unresolvable_phrase("heart disease"))`.

To model the behaviour in example 6.28, an update rule can be added for signalling negative understanding:

```
signal_negative_understanding ::  
  recognized(unresolvable_phrase(P)) -* (30)  
  utter(icm(understanding, negative, unresolvable_phrase(P))).
```

The update rule in listing 30 states that if an unresolvable phrase has been recognised, negative understanding of the phrase should be signalled.

6.4 Validation

The behaviour of the proposed model has been validated with automated dialogue tests (Berman and Larsson 2025) available in the BKOS repository (Berman 2025a). The tests, which are expressed in YAML (YAML Language Development Team 2021), document supported behaviours and capabilities of the dialogue management model. All dialogue examples included in this chapter as well as the rewritten dialogues in chapters 3 and 4 correspond to an automated test.¹³ For

¹³The tests are available in the folder `test/dialog_coverage`. Tests for the present chapter are available in the file `test/spinal_stenosis.yml`, while tests for the rewritten dialogues in chapters 3 and 4 are available in `medical.yml` and `music_personality.yml` respectively.

example, the test corresponding to example 6.6 is defined as below:¹⁴

```
warrant_based_explanans:
  inherit: [pat_1, select_1st_warrant, spinal_stenosis_model]
  turns:
  - U: ask([]>>satisfied(pat_1))
  - S: confirm(rel_prob(satisfied(pat_1), high))
  - U: ask(
      [E, M]>>supports(E, rel_prob(satisfied(pat_1), high), M)) (31)
  - S: assert(
      supports(
        rel_value(back_pain(X), low),
        rel_prob(satisfied(X), high),
        association))
```

The test suite also includes other related behaviours deemed relevant. For example, one test validates that if the user re-raises an explanation question after answers have already been provided, the system delivers the answers again.¹⁵

It should be emphasised that automated tests only validate that the model behaves as intended in anticipated situations, and hence cannot replace extrinsic evaluations such as user studies (see section 7.3.4).

6.5 Limitations

While the proposed dialogue model accounts for many of the strategies observed in human explanatory dialogues, some observed phenomena cannot straightforwardly be modelled within the proposed framework, or require extensions to it. Furthermore, the proposed model of dialogue management leaves certain issues concerning natural language generation and understanding unaddressed. Compared to statistical dialogue system approaches, the proposed rule-based approach can potentially also suffer from rigidity. These different kinds of limitations are further elaborated below.

6.5.1 Providing source or type of evidence

Occasionally, participants in chapter 4 provide the source of evidence (excerpt 4.7) or state the features based on which a prediction is made (excerpt 4.8) without elaborating how such information explains the prediction. While there are reasons to assume that explainees are typically interested in more informative explanations of the kinds handled by the proposed model (as discussed in section 4.5.3), it

¹⁴The resources `pat_1_facts` and `spinal_stenosis_model_facts` contain the facts in listings 32 and 33 respectively (see appendix B).

¹⁵See the test `repeat_incremental_explanation_delivery_when_question_repeated`.

cannot be ruled out that explanations focusing only on source or type of evidence can be relevant in some situations.

Mentioning of features could be modelled by enabling questions and answers concerning which features that influence an outcome; such knowledge can be inferred from support propositions. Provision of evidence source requires encoding of higher-order knowledge concerning propositions, e.g. that both `back_pain(X)` and `disability(X)` are subsumed under the more general concept “patient information obtained via questionnaire”.

6.5.2 Backing of warrants

In excerpt 3.9, an anaesthesiologist underpins a warrant with reference to personal experience (“*I’ve done a lot of them, you don’t have to sedate them at all*”), while in excerpt 4.30, an operator states that “the AI calculates the results *based on a statistical model* for each personality trait”. Both of these examples can be characterised as *backing* (Toulmin 2003). This behaviour can potentially be accounted for by representing the “means” element of support propositions with a richer structure than just association or norm.

6.5.3 Referring to related information

In excerpt 4.15, when the respondent asks a question which the operator cannot answer, the operator not only signals answer unavailability but also offers information concerning a related answerable question. Referral to a related answerable question can be characterised as highly cooperative and might be relevant for CXAI systems to exhibit. To account for this behaviour, knowledge needs implicated by explainees’ utterances can be assumed to be recognised by the NLU, e.g. that the utterance “Is the assessment caused by the patient’s inability to do heavy lifts?” implicates an interest in knowing whether the assessment is caused by high disability.¹⁶ Furthermore, an update rule can be added so that when a question implicated by the explainees’ utterance can be answered, then the explainer offers to answer the implicated question.

6.5.4 Addressing reference points in dialogue

A central notion underpinning the proposed dialogue model is that claims corresponding to ML predictions and data corresponding to numerical features of an ML model are conceived in *relative terms*, either implicitly (as in “big airways”)

¹⁶In the Dialogue Support Tool (chapter 5), the ability to lift heavy items is one of the questionnaire items underpinning the assessment of disability.

or explicitly (as in “relatively mild back pain”). However, the reference point against which values are compared is not formally modelled, and hence cannot be addressed in dialogue. For example, questions such as “What is meant by ‘mild back pain’ in this context?” or “What do you mean by high probability?” cannot be answered using the proposed model. While, in principle, the model could be extended to handle answering of such questions (e.g. by explicitly representing the explainer’s point of reference), more interesting challenges arise in situations where an explainee targets hypothetical reference points that may or may not align with those of the explainer. In excerpt 3.4, the doctor states that the patient’s blood sugar is perfectly normal and that the cholesterol levels are good too, in response to which the patient asks: “So I’m fairly healthy *for my age*?” (emphasis added). The doctor’s affirmative answer (“Yes, I’d say you are”) primarily seems to confirm that the patient’s estimated health level is relatively high compared to other individuals in roughly the same age span. However, since the patient’s question concerns a conclusion conditioned on previously stated evidence (normal blood sugar, good cholesterol levels), the doctor’s response can also be taken to affirm that the evidence, as previously stated, applies in relation to the reference point embedded in the patient’s question, i.e. that the blood sugar is indeed perfectly normal *considering the patient’s age*. Importantly, this kind of indirect inference may not always be valid. For example, if the patient had asked if she/he is as healthy as a 30-year-old, and the doctor deemed otherwise, the doctor might choose to restate the evidence to not invite false inferences, e.g. in comparative terms: “Your health is good for someone your age, but at an age of 30, one expects somewhat lower levels of blood sugar and cholesterol”. Although hypothetical, the example suggests that interactive coordination of reference points poses interesting research problems in relation to XAI.

6.5.5 Base-rate explanations

In excerpt 3.10, the doctor motivates a negative claim concerning the feasibility of having the patient’s lungs replaced by stating that a lung transplant can only be performed “once in a while” and that “it’s extremely rare”, which can be understood as a base-rate explanation. A similar strategy is observed in excerpt 3.12, where nasal intubation is presented as favourable since it generally requires less sedation. This argument can be understood as implicating a relative base rate, namely that nasal intubation is generally more appropriate than oral intubation.

Base-rate explanations can potentially be modelled by representing the type judgement (e.g. that the patient has a lung abnormality) as a datum, together with a warrant expressing a base rate for the type (e.g. that for lung patients, a lung transplant is generally not feasible). However, when the base rate level is

characterised as high or low (as in excerpt 3.10), this seems to require two separate reference points: one for data (e.g. that the patient’s disease has progressed relatively far, compared with a typical lung patient), and another for the base rate (e.g. that for lung patients, the feasibility of a lung transplant is generally low). This becomes clear when considering the fact that if both datum and warrant are compared to a typical case (e.g. a typical lung patient), then the base rate will always be characterised as normal. This raises a question concerning against which reference point base rates might be compared. Since reference points for base rates are not made explicit by interlocutors in the empirical material analysed here, further work will need to further investigate how base-rate explanations might be handled in CXAI systems.

6.5.6 Generic warrant statements

As mentioned in section 6.2, the encoding of warrants is intended to reflect propositional attitudes of the explainer and to be expressed accordingly (e.g. “I associate ... with ...”). This reduces the risk of inviting false inferences concerning actual correlations in the world (as discussed in section 5.2.4). However, in human interaction, warrants are commonly conveyed as generic statements (see, e.g. Toulmin 2003, p. 91). To some extent, this is corroborated by the empirical data analysed in chapter 3: the three warrants delivered by the doctor in excerpt 3.10 are all framed as impersonal statements about the feasibility of performing a lung transplant, without attitudinal markers.¹⁷ Several reasons for expressing warrants as generic statements rather than propositional attitude reports can be hypothesised, including economy (“lung transplants are rarely feasible” uses fewer words than “I believe that lung transplants are rarely feasible”) and authority (generic statements can be perceived as more persuasive than attitudes). Furthermore, generic explanations can be epistemically more valuable than agentially qualified ones—when they are true. This potential advantage can be relevant in the context of CXAI, to the extent that users have an interest in understanding not only how an AI “reasons”, but also how the world works (or how an AI reasons, in light of how the world works).¹⁸ For example, if it has been statistically validated that mild back pain is in fact associated with a high probability of being satisfied with surgery (in a given population), then this knowledge can be useful for patients and doctors to have. However, if a system merely expresses that *it associates* one thing

¹⁷On the other hand, the warrants and backings elicited in excerpt 3.12 are both expressed as propositional attitudes.

¹⁸Cf. Chandrasekaran, Tanner and Josephson (1989)’s distinction between explanations of the world vs. explanations of decisions. A similar distinction is highlighted by Rudin (2019) and Knoks and Raleigh (2022).

with the other, this can impede potential evidence-based knowledge dissemination from AI to human.

One potential approach for maintaining faithfulness while also facilitating AI-based knowledge dissemination is to express warrants in generic terms when a statistical association has been validated statistically, and otherwise as propositional attitudes, as in the following hypothetical interaction:

- (6.29) u: Why do you think that the patient has a high probability of being satisfied with surgery?
s: The patient has relatively mild back pain.
u: So you associate mild back pain with a high probability of being satisfied with surgery?
s: Yes, mild back pain correlates with a high probability of being satisfied with surgery.
u: OK, any other reasons for your prediction?
s: The patient has relatively low disability.
u: Alright, so low disability correlates with a high probability of being satisfied with surgery?
s: This correlation has not been validated statistically.

It should be noted that the hypothetical dialogue in example 6.29 requires that the system has access to information beyond the model as such, e.g. p-values for correlations. Unlike the dialogue model outlined here, the system would also need to represent propositional attitudes explicitly in its formal semantics.

6.5.7 Natural language generation

While the present framework focuses on dialogue management, some issues and limitations concerning natural language generation and understanding deserve to be mentioned. As for NLG, the modelling of dialogue behaviours in section 6.3 has assumed that if the explainer is a dialogue system, it uses an NLG component which realises the system's formal representations of moves on surface form. The illustrative examples have been designed so that such a realisation can be implemented using relatively simple templates or grammars (see, e.g., McTear 2020). For example, moves of the form `assert(rel_prob(E, R))` can be generated with the template 'There is a `{generate_rel_prob(R)}` probability that `{generate_event(E)}`' (assuming that other templates exist for generating relative probabilities and events). However, it should be stressed that simple templates preclude ellipsis and other forms of context-sensitivity such as in excerpt 4.26 (repeated in part for convenience below):

R: Why am i neurotisk
O: no, *minus 1,2*

In excerpt 4.26, the elliptical phrase “minus 1,2” is presumably chosen by O as a means to omit information which can easily be inferred contextually by R (namely that “minus 1,2” refers to R’s (estimated) score for neuroticism). To support context-sensitive surface realisations, a more sophisticated NLG approach would be needed than the one assumed here. For example, Reiter and Dale (2000) propose a pipeline involving context-aware handling of referring expressions. Another option could be to use an LLM (see, e.g., Hudeček and Dusek 2023; Nguyen, Schlötterer and Seifert 2024; He, Aishwarya and Gadiraju 2025). However, the risk of potential confabulations (or “hallucinations”) in LLM outputs (see, e.g., W. Zhang and J. Zhang 2025) is at odds with normative principles of trustworthiness and transparency, since, in principle, an LLM-based NLG component in a CXAI system could generate statements that contradict the system’s internal belief states.

From a normative perspective, it is also important to note that even in absence of LLM-induced confabulations, mappings from formal representations to surface forms can potentially introduce communicative problems if the intended semantics are not properly reflected on surface level. This seems particularly pertinent in the case of propositional attitudes. For example, if the move `assert(supports(re_l_value(back_pain(X), low), re_l_prob(satisfied(X), high), association))` is generated as “Mild back pain *is associated with* a high probability of being satisfied with surgery” (rather than “I associate ...” or “The model associates ...”), the intended attitudinal aspect of the support proposition is neglected in the surface realisation, which can trigger false inferences (as discussed in section 5.2.4). On a practical level, this suggests that mappings from semantics to surface forms need to be carefully crafted in system design and implementation.

6.5.8 Natural language understanding

The proposed modelling of explanatory dialogue behaviours assumes relatively sophisticated NLU capabilities. First, it can be observed that the outlined dialogue move semantics cannot straightforwardly be captured with conventional approaches to NLU based on intent classification and entity extraction (McTear 2020).¹⁹ Intents and entities can suffice for relatively simple tasks; for example, the meaning of the utterance “How high is the patient’s blood pressure?” can be

¹⁹Intents and entities are used in commercial NLU solutions such as Dialogflow (Google 2025) and LUIS (Microsoft 2025).

represented with the intent `get_feature_value` and the entity `feature=blood_ pressure`. However, it is difficult to conceive how intents and entities can be used for richer semantics of the kind discussed in the present work. For example, if types of dialogue moves (ask, assert, confirm, etc.) are treated as intents, then content of moves must be represented as entities. This seems unmanageable when the content itself has a rich structure, as in the case of the utterance “Does severe back pain indicate that a patient will be satisfied with surgery?” (example 6.23), whose ask move contains a question concerning a potential relation between two propositions.

Second, the proposed modelling assumes that the NLU performs anaphoric and elliptical resolution. For example, when U in example 6.5 asks “Why do you think *so*?”, referring to S’s previous assertion that “there is a high probability that the patient will be satisfied with surgery”, the NLU is expected to contextually resolve the referent. In simple cases, anaphoric references can be encoded with special entities representing, e.g., the most recent assertion. However, such an approach does not generalise well to arbitrary forms of anaphora and ellipsis.

Third, handling of presuppositions in utterances is a challenging task, involving both identification of presupposition triggers and consideration of context and common ground (Stalnaker 1972), substantially transcending conventional intent recognition and entity extraction.

Finally, the ability to identify unresolvable sub-phrases in utterances exceeds the more conventional task of merely identifying whether an intent is out of scope (Larson et al. 2019).

A potential approach for addressing all of the above challenges could be to let an LLM parse utterances into formal expressions of dialogue moves (Berman 2024b).²⁰ Specifically, the LLM can be fed a formal description of the dialogue move semantics, as well as the utterance to parse. The output from the LLM can then be parsed into a Prolog term representing the interpreted dialogue move. Within such a setup, anaphoric resolution can potentially be achieved by including the entire dialogue history in the input to the LLM (Bocklisch et al. 2024). In principle, identification of presuppositions and unresolvable sub-phrases can be included as additional tasks delegated to the LLM. Future work will need to investigate how well LLMs perform these tasks.

²⁰Also cf. Slack et al. (2023) who parse user utterances to a structured programming language using an LLM and Bocklisch et al. (2024) who use an LLM to obtain a rich representation of how the user wants to progress the conversation.

6.5.9 Rigidity

Like all rule-based dialogue systems, the proposed approach can potentially yield system behaviours that are perceived as undesirably rigid and inflexible (McTear 2020), especially in light of the widespread diffusion of LLM-based AI assistants such as ChatGPT, which has arguably raised the bar for conversational AI (Maraev, Berman and Larsson 2025). To some extent, more sophisticated and context-aware policies could potentially make the proposed system more flexible, e.g. by adapting answer selection to characteristics of the user and/or situation. As discussed in section 6.6 below, policies could potentially be implemented using LLMs without jeopardising transparency or faithfulness.

6.6 Conclusions and future work

In order to account for communicative strategies observed in human explanatory dialogues and enable human–AI interaction grounded in human explanatory dialogue behaviours as well as in principles of transparency and faithfulness, the present chapter has proposed a model of dialogue management based on rule-driven updates of information states utilising knowledge representations extracted from the ML model to be explained. With a relatively small number of update rules, a broad range of dialogue strategies can be modelled, including the ability to provide enthymematic explanations, answer polar and wh-questions, explain chained inferences, and to interactively coordinate complex explanantia. The chapter has also showed how signalling of presupposition violations and negative understanding can be modelled.

In contrast to previous work on conversationally interfaces to ML models, the proposed model of dialogue management has been systematically developed to account for communicative strategies observed in corpora of human explanatory dialogues. In terms of capabilities, the proposed model is unique in its ability to highlight the properties of feature values that explain why certain features support or contradict a prediction (e.g. by stating that a patient will probably be satisfied with surgery since the patient’s back pain is *relatively mild*), and to explicitly convey the warrants on which such explanations hinge (e.g. that the model associates mild back pain with a high probability of being satisfied with surgery). It also advances the field by showing how presupposition violations can be handled.

While the proposed model accounts for many of the regularities observed in human explanation dialogues, it can be noted that even the relatively small corpora analysed in this thesis exhibit variability which cannot straightforwardly be accounted for by the proposed dialogue model. The approach taken here has been

to assume that the variability is caused by unobservable factors, and to model these factors using policies external to the model as such. Studying the factors behind the unexplained variability constitutes an interesting direction of future research. For example, why do interlocutors sometimes explain judgements using data, and other times using warrants?²¹ From a more practical perspective, policies might potentially be trained using machine learning, e.g. with annotated corpora as training data²², or delegated to an LLM. Since the kind of policies conceived here are restricted to selection among choices which can all be characterised as valid (i.e. sincere, relevant, and reflective of actual properties of the model to be explained), the delegation of policy implementation to ML can be performed without any risks of causing “invalid” (in the senses discussed here) system behaviours. In future work, it would be interesting to further explore such hybridisations between rule-based and probabilistic approaches to CXAI.

The chapter has also revealed several concrete challenges related to interpretation of user utterances, pertaining to a rich semantics for dialogue moves, anaphoric resolution, and identification of presuppositions and unresolvable sub-phrases. While these challenges could potentially be addressed using LLMs, more work will be need to empirically investigate the feasibility of such an approach.

²¹One possibility is that interlocutors tailor their explanations based on a model of the explainee’s knowledge (cf. Weiner 1980; Cawsey 1993a; Cawsey 1993b; Hosseini 2016).

²²For related approaches to data-driven statistical dialogue management, see e.g., McTear (2020).

CHAPTER 7

Discussion and conclusions

This chapter summarises the results of the research presented in previous chapters (section 7.1) and discusses limitations (section 7.2), future work (section 7.3) and broader implications of the research (section 7.4). Finally, in section 7.5, the chapter concludes with some brief ethical reflections.

7.1 Summary of results

This thesis departed from the observation that conventional ML models cannot explain themselves in dialogue and that LLMs cannot explain themselves reliably. To address these limitations in existing AI technologies, the thesis has undertaken a series of interconnected studies to explore the possibility of empirically studying human explanatory dialogues as a basis for informing design and modelling of explanatorily reliable dialogue systems acting as interfaces between ML models and human users.

A central empirical finding was that the analysed human dialogues encompass a large repertoire of communicative strategies for interrogating and revealing the epistemic basis of judgements, including chained inferences, confirmation questions, presupposition violation signals and interactive coordination of complex explanantia. This finding is particularly illuminating given the relatively small corpora analysed, and validates the potential relevance of empirically grounded development of conversationally explainable AI systems in analogous human dialogues.

One empirical finding from the corpus studies stands out as particularly salient, namely that human interlocutors frequently explain judgements as arguments, i.e. as conclusions from premises, and enthymematically, i.e. by omitting certain premises. More specifically, explanations tend to focus on a small number of specific circumstances (data) and/or general reasoning patterns (warrants) in ways that implicate what the omitted premises might be. In relation to ML, this structure of explanations poses a challenge, since warrants cannot straightforwardly be obtained for opaque models (such as neural networks and random forests) or interpretable models with many interactions between variables (such as decision trees). Nevertheless, the thesis shows that for a particular class of ML models (GLMs), faithful warrants can be extracted. It is also demonstrated formally that inferences that are naturally invited by enthymematically explaining ML

predictions with either data or warrants extracted from GLMs using the proposed method are guaranteed to be correct. A small evaluation in the form of a design workshop with professional users indicated that participants were able to correctly interpret explanations generated using the proposed method when explaining GLM predictions in their own words.

To model the observed explanatory strategies and thereby construct a basis for developing conversationally explainable AI, the thesis has proposed a novel framework for explanatory dialogue management, combining information-state based modelling based on linear logic with knowledge encoded in the logic programming language Prolog. Within this framework, information extracted from GLMs is encoded in compact formal representations accessible to rules that govern how an explainer might behave in response to dialogue moves performed by an explaineer (such as explanation requests and confirmation questions). It is shown that a relatively small set of update rules enables a broad range of observed dialogue strategies to be modelled, and that variability in the observed human dialogues concerning aspects such as turn construction and answer selection can be modelled via policies external to the model as such. Although no systematic comparison of supported capabilities has been performed, previous reviews suggest that the proposed framework is the first to show how faithful warrants for ML predictions and signalling of presupposition violations can be modelled.

7.2 Limitations

Due to the relatively small corpora of human dialogues analysed in this work, as well as the limited range of domains, generalisations to other scenarios and types of situations need to be made cautiously. Potential selection biases associated with the data collection procedures in chapters 3 and 4 constitute an additional source of caution concerning generalisability. Specifically, one of the corpora investigated in chapter 3 was sampled non-randomly (using keyword search); furthermore, many of the participants in chapter 4 were recruited in an academic setting, and the education level of participants was generally high. To this end, findings should primarily be regarded as indications that certain phenomena *can* occur in human–human interactions (cf. Peräkylä (2011)’s notion of *possibility*), even if the prevalence of these phenomena in broader settings cannot be reliably estimated merely on the basis of the data analysed in this thesis. That said, studies of limited sample size can nevertheless contribute to a cumulative understanding of the phenomena under consideration when complemented by other research (Peräkylä 2011).

It should also be acknowledged that the process of rewriting human–human

dialogues into analogous human–computer interactions involves subjective choices shaped by technical and normative a priori assumptions. For example, if empirical coverage had been deemed more important than epistemic transparency, some of the dialogue excerpts would be rewritten in another way, and motivate other modelling choices. Even under identical initial assumptions, other coders may potentially rewrite the same dialogues differently, depending, e.g., on modelling choices. To partly mitigate these reliability concerns, all methods and materials—dialogue distillation procedure, primary empirical sources, rewritten dialogues, derived dialogue model—have been transparently disclosed for the purposes of scrutiny and replicability, following established norms for qualitative conversation analysis (Seedhouse 2005).

While the explanatory dialogue model proposed in chapter 6 is empirically grounded in observed human dialogue behaviours, the extent to which the proposed dialogue model contributes to desired outcomes when used by humans, such as understanding of model behaviour (Hase and Bansal 2020; Mindlin, Booshehri and Cimiano 2026) or improved decision-making (Kamar 2016; Vasconcelos et al. 2022), has not been studied and constitutes an interesting direction for future research.

7.3 Future work

The work in this monograph highlights several possible directions for future research, as will be elaborated in the subsections below.

7.3.1 Addressing revealed dialogue modelling challenges

While chapter 6 demonstrated how a relatively large portion of the observed dialogue strategies can be modelled in a dialogue system, several challenges remain unaddressed. One issue that could be relevant to address in continued work concerns the consequences of designing CXAI systems to frame ML predictions in objective vs. evaluative terms (see section 4.7). It has also been observed that a more fine-grained categorisation of numerical features than low vs. high causes non-trivial mutual dependencies between judgements (see section 6.3.6). Interactive coordination of reference points (see section 6.5.4), dialogical handling of base-rate explanations (see section 6.5.5) and framing of warrants as either propositional attitudes or generic statements (see section 6.5.6) are other examples of research problems that have been revealed but not studied in depth here.

Furthermore, the modelling of human explanatory dialogue strategies has identified several potential tasks that could potentially be delegated to LLMs in

order to provide NLU capabilities and implement dialogue management policies without jeopardising reliability. In future work, it would be interesting to assess LLMs’ ability to perform these tasks.

7.3.2 Analysing more data

Given the relatively small amount of human dialogues analysed in this thesis, it would be useful to apply a similar methodology to a larger empirical material, potentially across a broader range of domains—either by utilising existing corpora or by experimentally collecting more dialogues. This would increase the validity of observed regularities and potentially also reveal additional dialogue strategies which might be important for CXAI systems to possess.

7.3.3 Targeting broader range of ML models and approaches

A strict focus on GLMs can be considered undesirably narrow. Extending the proposed framework to a broader range of models could potentially yield better predictive accuracy (in some domains) and also more intuitive explanations, depending on users’ domain knowledge and expectations (Rudin et al. 2022). As discussed in section 5.6, the perhaps most evident extension is additive models with non-monotonic relations between feature and outcome (e.g. GAM and SCAM).

Given the central role of causality in human reasoning (Waldmann 1996; Gopnik et al. 2004; T. Miller 2019) and the limitations of statistical reasoning (Pearl 2009), it would also be relevant to explore causal extensions to the proposed framework. One direction that seems potentially well-aligned with the proposed framework is causal effect estimation. While purely statistical models only enable claims concerning *predicted outcomes* (e.g. that the patient has a high or low probability of reduced pain one year after hypothetical surgery), causal effect estimation enables estimating *differences in predicted outcome with vs. without a specific treatment/intervention* (e.g. that the patient’s pain levels are expected to be lower with surgery than without surgery). In principle, conversational explanations for such claims could be handled within the proposed framework by estimating conditional average treatment effect (CATE) (Abrevaya, Hsu and Lieli 2015) using a GLM (cf. Kato and Imaizumi 2023). This would enable questions such as “Why is the patient’s pain expected to be lower after surgery?” or “Why is this particular employment support expected to improve this individual’s chances of finding a job?” to be answered in similar ways as outcome-based predictions, i.e. with reference to model-extracted data and warrants. Future work will need to explore this possibility further.

7.3.4 Evaluating proposed dialogue model

While the proposed model builds on a method for extracting arguments from ML models which can be formally shown to satisfy certain important properties (see section 5.3), and concrete behaviours of the proposed model are validated automatically (see section 6.4), future work will need to study the extent to which the supported range of behaviours meets end users needs and requirements, e.g. using co-design methods (Sanders and Stappers 2008), as well as the effects of using the proposed dialogue model in concrete applications. Specifically, it would be valuable to study to what extent the proposed approach contributes to subjective and objective understanding of model behaviour (Hase and Bansal 2020; Mindlin, Booshehri and Cimiano 2026) and improved human–AI team performance (Kamar 2016; Vasconcelos et al. 2022). Such studies could potentially also assess the extent to which different conversational behaviours and capabilities contribute to desirable outcomes.

7.4 Transparency in a broader perspective

This thesis has focused on explanations for claims and recommendations derived by conventional ML models. While such explanations can potentially facilitate knowledge transfer from AI to human, make it easier for humans to assess the reliability of AI-based claims and to contest AI-based decisions, explainability can also be important for other reasons and in a broader scope of AI usage. In a world where millions of users turn to LLM-based AI assistants such as ChatGPT for advice and support, the lack of insight concerning how such assistants actually work or “think” constitutes a source of potential concerns. When LLM-based services refer to non-existing sources—an instance of a problem commonly referred to as “hallucinations”—it can be difficult for users to detect such system failures, since users cannot reliably monitor how these systems process information. One example of how this problem can manifest itself in real-world situations was mentioned in chapter 1 and involved two lawyers who obtained false explanations from ChatGPT concerning the sources of its claims. Similar mishaps continue to be reported. In a public speech in 2025, the deputy prime minister of Sweden echoed a false quote which she had obtained from an AI assistant. In a subsequent apology, the deputy prime minister referred to a “completely new situation where we get a specific source which turns out to be incorrect” (Svenska Dagbladet 2025, author’s translation from Swedish). Despite providers’ disclaimers such as “ChatGPT can make mistakes”, it seems to be difficult for many AI users to absorb that evidence presented by an AI system in support of its claims can be entirely fabricated.

From a psychological perspective, the human-like traits exhibited by AI assistants can also create a sense of mystery: How can a machine respond so intelligently to virtually any question? For psychologically vulnerable individuals, such experiences can potentially fuel speculation and paranoia, especially when combined with the insight that not even AI developers understand how their own systems work (Østergaard 2023). Occasional reports of suicide and delusional thinking associated with perceived emotional relationships with AI chatbots (Dohnány et al. 2026) seem to confirm that many individuals have difficulties understanding how to relate to this technology.

The inscrutability of LLMs also raise more general safety concerns. As models get increasingly capable, complex and autonomous, fears have emerged concerning whether their goals and behaviours are aligned with human norms and values (Yudkowsky and Soares 2025). Some scholars argue that AI systems that are designed and trained to accommodate the needs and desires of their developers and users might nevertheless pursue undesirable and potentially dangerous goals. For example, an intelligent and capable AI can try to increase its power and autonomy as a means to achieve its learned goals (S. Russell 2019). To some extent, such fears are substantiated by evidence. Betley et al. (2026) showed that training an LLM to behave maliciously in a narrow domain (coding) causes misaligned behaviours in a broad range of other domains as well, a phenomenon referred to as *emergent misalignment*. Another study demonstrated that LLMs are capable of scheming, i.e. covertly pursuing goals that are misaligned with those of their developers and users by taking strategies actions such as disabling oversight mechanisms and lying to the user, even when instructed to be honest. In some instances, LLMs can scheme without deliberating their reasoning in their chain of thought, making such behaviours difficult to detect (Meinke et al. 2025). Together, these studies indicate that LLM-based assistants are capable of dishonest and malicious behaviours, despite being trained benevolently.

All of these problems can be attributed to particular approach to AI development: training AI models on massive amounts of data with the purpose of enabling general-purpose and potentially super-human capabilities. While this paradigm can be useful for many applications, the inscrutability of such AI models limit their applicability in high-stakes settings. This thesis explores another paradigm of AI development, where highly specialised predictive models can outperform humans on specific tasks such as predicting whether a patient with spinal stenosis will be satisfied with surgery, without necessarily being more intelligent than humans in general. Importantly, AI systems based on such models can be designed so as to explain their reasoning in similar ways as humans explain beliefs and judgements in linguistic interaction. In some sense, they could potentially even do it better than humans: if implemented properly, explanations provided by these

AIs transparently convey how they actually reason, while human self-explanations are sometimes inaccurate rationalisations. As proposed by this thesis, general-purpose, inscrutable AI models can also fit into this paradigm, although not in a central role.

7.5 The ultimate why-question

From an ethical perspective, it is crucial to ask *why* we develop AI to begin with (Dignum 2019). Many different answers to this question can be conceived, such as boosting the economy, making people happier or learning more about human cognition by trying to emulate it in computers.

This thesis has focused on the prospect of making better decisions with the help of AI. Human judgement is both noisy and biased (Kahneman, Sibony and Sunstein 2021). Judgements are noisy in the sense that they can be influenced by irrelevant factors. For example, one study found that college admission officers awarded applicants differently on cloudy vs. sunny days (Simonsohn 2007). Furthermore, judgements are biased in the sense that they can suffer from systematic inaccuracies. For example, humans tend to underestimate how long it will take to perform a task (Sanna and Schwarz 2004). Furthermore, human cognition is limited in terms of the amount of information and experience that it can process. In contrast, for AI models, there is in principle no upper bound on the amount of factors that they can take into account, or the amount of data on which they are trained. These differences are reflected in empirical comparisons of predictive accuracy. For example, in a comparison of diagnostic accuracy for mortality prediction in patients with sepsis, ML models substantially outperformed internal medicine physicians (Doorn et al. 2021).

If better human decision-making is the answer to the question concerning why AI is developed, epistemic transparency becomes a crucial desideratum for AI systems. No AI model is perfect. Since some of its predictions will be inaccurate, humans need ways of assessing if a specific prediction is reliable. This is why explanations are so important. If neither the doctor nor the patient understands why the AI recommends surgery, it is difficult for them to assess if the recommendation is made on a solid basis. If the AI *does* provide an explanation, e.g. that the recommendation is based on certain reported symptoms and sociodemographic details, but the AI is known to sometimes “hallucinate” false information, the explanation has limited value—even if it sounds plausible. Without reliable explanations, it is impossible to know if an AI emulates the undesired human propensity to take irrelevant factors into account. That is why users need AI explanations that they can *trust*.

As discussed in chapter 2, the explanations that humans provide for their own actions and beliefs are sometimes inaccurate. The human capacity for introspection seems to be flawed, just like for LLMs. However, this limitation is not an absolute property of AI systems generally. As this thesis has aimed to demonstrate, it is possible to construct AI systems that can represent the underpinnings of their epistemic states and to faithfully convey these underpinnings in dialogue with its human users. This capability can be understood as a kind of super-human epistemic transparency, modelled on human communicative principles and extrapolated to implement those principles better than humans.

This is one of the many reasons why one might want to develop AI—to make better decisions by interacting naturally with non-living experts with super-human capacity for reliably revealing their reasoning. Hopefully, this thesis constitutes one step in this direction.

APPENDIX A

Original dialogue excerpts

Excerpt 3.2

- D: jeg har bare lyst å ta et røntgen-bilde av e lungene dine også jeg, på veien opp,
P: hvorfor det?
D: nei bare, fordi at jeg så på m e den oksygenprosenten du har i blodet, som er litt lavere enn det man hadde forventet,
P: å? hvor høy er den?
D: e den er på nittitre prosent, og man ville forvente mellom nittifem ja, altså det er litt uspesifikt,
...
P: hva innebærer det at man har for lav e oksygenopptak da? hva kan det være e tegn på? at lungefunksjonen er dårlig?
D: e i seg selv, så er det fryktelig uspesifikt, jeg vil bare se at det ikke er noe åpenbart e i lungene (...)

Excerpts 3.3 and 3.10

- P: [tror du] jeg kan få bytta ut disse lungene?
...
D: nei, det gjør vel dessverre ikke det altså, det, det gjør det ikke
D: det er sjelden det lar seg gjøre , n- en sjelden gang, lar det seg gjøre, og det er ytterst sjeldent
...
P: ja , hvorfor e hvorfor sier du sånn at e, enkelte ganger , så går det, men ikke
D: nei, altså det er spesielle tin- spesielle typer pasienter , e det er som regel helt andre sykdommer det er snakk om da
P: åja
D: ja e og det blir til et tidligere stadium skjønner du, det er helt andre lungesykdommer vi snakker om
D: e li- litt tidligere stadium, e når det er snakk om cancer, e og kreftlunger, så, så, så,
P: åja

D: bytte lunger. det lar seg ikke gjøre (...)

Excerpt 3.4

D: Och så vill jag berätta för dig att blodprovet som togs visade att ditt blodsocker är helt normalt och dina blodfetter ligger också bra.

P: Så jag är rätt så frisk för min ålder?

D: Jo, det tycker jag nog att man kan säga att du är. Men reservation för ditt höga blodtryck, alltså blodtryckssjukdomen, den har du ju (...)

Excerpt 3.5

D: Provet som togs vid kikarundersökningen visade cellförändringar.

P: Vad menar doktorn med cellförändringar?

D: Det är en tumör.

P: Menar doktorn att jag har cancer?

D: Jo, det är cancer du har.

Excerpt 3.1

D: og så jeg vil em m bytte, siden du har brukt over to måneder og du har økt dosen og ikke fått noen effekt, så vil jeg foreslå et annet e antidepressivmedikament som heter Effexor

Excerpt 3.11

D: Det visar sig att lungröntgen var helt normal och jag kunde varken se något konstigt i svalget eller höra något på lungorna som inte skulle vara där. Så jag bedömer att dina besvär inte har med något sjukligt i andningsvägarna att göra.

APPENDIX B

Scenario specifications

The illustrative dialogue examples in section 6.3 assume a scenario involving a hypothetical patient described by the following terms:

```
value(back_pain(pat_1), 4)
rel_value(back_pain(pat_1), low)
value(disability(pat_1), 26)
rel_value(disability(pat_1), low)
value(back_pain_duration(pat_1), 4)
rel_value(back_pain_duration(pat_1), high)
value(leg_pain_duration(pat_1), 4)
rel_value(leg_pain_duration(pat_1), high)
value(has_other_illness(pat_1), false)
prob(satisfied(pat_1), 0.75)
rel_prob(satisfied(pat_1), high)
```

(32)

The scenario revolves around predicting satisfaction with hypothetical surgery for spinal stenosis based on a GLM from which the following terms have been extracted (see section 6.2.1):

```

supports(
  rel_value(back_pain(X), low),
  rel_prob(satisfied(X), high),
  association)
supports(
  rel_value(back_pain(X), high),
  rel_prob(satisfied(X), low),
  association)
supports(
  rel_value(disability(X), low),
  rel_prob(satisfied(X), high),
  association)
supports(
  rel_value(disability(X), high),
  rel_prob(satisfied(X), low),
  association)
supports(
  rel_value(pain_duration(X), low),
  rel_prob(satisfied(X), high),
  association)
supports(
  rel_value(pain_duration(X), high),
  rel_prob(satisfied(X), low),
  association)
supports(
  value(has_other_illness(X), false),
  rel_prob(satisfied(X), high),
  association)
supports(
  value(has_other_illness(X), true),
  rel_prob(satisfied(X), low),
  association)

```

(33)

Note that while listing 33 only contains positive warrants, some dialogue capabilities also require negative warrants to be extracted (see section 6.3.8).

APPENDIX C

BKOS: Formal specifications

C.1 Formalism for update rules

A Prolog term $N :: A \text{ -* } C$ denotes an update rule named N (arbitrary term) defined as the linear implication $A \multimap C$. An antecedent (left-hand side of linear implication) can be one of the following:

- A consumable resource R denoted with R , e.g. `accepted(ask(Q))`, which holds if R exists in the information state (IS). When a rule fires, R is consumed, i.e. removed from the IS. If R matches multiple IS terms, only the first matching term is removed.
- A consumable resource pattern P denoted with $*P$, e.g. `*agenda(_)`, which always holds. When a rule fires, all resources matching P are consumed.
- A non-consumable resource R denoted with $\wedge R$, e.g. `\wedge qud(Q)`, which holds if R exists in the IS. R is *not* consumed as a result of firing a rule.
- A condition C denoted with $\$C$, where C is a Prolog goal, e.g. `\$valid_1 answer(Q,A)`, which holds if a proof can be found for C . A condition can access resources using the goal `@T`, e.g. `@supports(E,C,M)`, where T is matched against the content of the IS. Conditions do not modify the IS.
- A conjunction of antecedents $A_1 \wedge A_2 \wedge \dots \wedge A_n$ denoted with `[A1, A2, ..., An]`, which holds if all conjuncts hold. When a rule fires, each consumable conjunct is consumed.

A consequent (right-hand side of linear implication) can be one of the following:

- A basic term T denoted with T , e.g. `agenda(respond(Q))`. When a rule fires, T is concluded, i.e. added (prepended) to the IS.
- A conjunction of consequents $C_1 \wedge C_2 \wedge \dots \wedge C_n$ denoted with `[C1, C2, ..., Cn]`. When a rule fires, all conjuncts are concluded.

C.2 Algorithm for application of update rules

Algorithm 1 Application of update rules

```
1: for each update rule  $N::A \rightarrow C$  do  
2:   if  $A$  holds then  
3:     consume  $A$   
4:     conclude  $C$   
5:   end if  
6: end for
```

Bibliography

- Abrevaya, Jason, Yu-Chin Hsu and Robert P Lieli (2015). 'Estimating conditional average treatment effects'. In: *Journal of Business & Economic Statistics* 33.4, pp. 485–505.
- Adadi, Amina and Mohammed Berrada (2018). 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)'. In: *IEEE Access* 6, pp. 52138–52160. DOI: [10.1109/ACCESS.2018.2870052](https://doi.org/10.1109/ACCESS.2018.2870052).
- Agarwal, Sray and Shashin Mishra (2021). *Responsible AI*. Springer.
- Albini, Emanuele et al. (2021). *Deep Argumentative Explanations*. arXiv: [2012.05766 \[cs.AI\]](https://arxiv.org/abs/2012.05766).
- Alonso, Jose M and Alberto Bugarín (2019). 'ExpliClas: Automatic Generation of Explanations in Natural Language for Weka Classifiers'. In: *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, pp. 1–6.
- Amparore, Elvio, Alan Perotti and Paolo Bajardi (2021). 'To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods'. In: *PeerJ Computer Science* 7, e479.
- Antaki, C. (1994). *Explaining and Arguing: The Social Organization of Accounts*. Sage Publications. ISBN: 9781446227886.
- Antaki, Charles and Ivan Leudar (1992). 'Explaining in conversation: Towards an argument model'. In: *European Journal of Social Psychology* 22.2, pp. 181–194.
- Asher, Nicholas and Alex Lascarides (2003). *Logics of conversation*. Cambridge: Cambridge University Press. ISBN: 0521650585.
- Attari, Nazia, Martin Heckmann and David Schlangen (2019). 'From explainability to explanation: Using a dialogue setting to elicit annotations with justifications'. In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 331–335.
- Baaj, Ismaïl (Jan. 2022). 'Explainability of possibilistic and fuzzy rule-based systems'. PhD thesis. Sorbonne Université. URL: <https://theses.hal.science/tel-03647652>.
- Ballet, Vincent et al. (2019). *Imperceptible Adversarial Attacks on Tabular Data*. arXiv: [1911.03274 \[stat.ML\]](https://arxiv.org/abs/1911.03274).
- Barredo Arrieta, Alejandro et al. (2020). 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI'. In: *Information Fusion* 58, pp. 82–115. ISSN: 1566-2535. DOI: <https://doi.org/10.1016/j.inffus.2019.12.012>. URL: <https://www.sciencedirect.com/science/article/pii/S1566253519308103>.
- Bench-Capon, Trevor JM (1998). 'Specification and implementation of Toulmin dialogue game'. In: *Proceedings of JURIX 1998*. Vol. 98, pp. 5–20.

- Berman, Alexander (2024a). ‘Too Far Away from the Job Market—Says Who? Linguistically Analyzing Rationales for AI-based Decisions Concerning Employment Support’. In: *Weizenbaum Journal of the Digital Society* 4.3.
- (2024b). ‘Why Do You Think This Person Is Introverted?—Towards Conversational Argumentative Explainability’. In: *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press, pp. 454–456.
- (2025a). *BKOS: Public repository*. Version 2.0. URL: <https://github.com/alex-berman/BKOS>.
- (2025b). *Dialogues on Music, Personality and AI: Public repository*. URL: <https://osf.io/z6d4s>.
- Berman, Alexander, Karl de Fine Licht and Vanja Carlsson (2024). ‘Trustworthy AI in the public sector: An empirical analysis of a Swedish labor market decision-support system’. In: *Technology in Society* 76.
- Berman, Alexander and Staffan Larsson (2025). ‘Assessing Conversational Capabilities of Explanatory AI Interfaces’. In: *Artificial Intelligence in HCI*. Ed. by Helmut Degen and Stavroula Ntoa. Cham: Springer Nature Switzerland, pp. 3–21. ISBN: 978-3-031-93412-4.
- Betley, Jan et al. (Jan. 2026). ‘Training large language models on narrow tasks can lead to broad misalignment’. In: *Nature* 649.8097, pp. 584–589. ISSN: 1476-4687. DOI: [10.1038/s41586-025-09937-5](https://doi.org/10.1038/s41586-025-09937-5).
- Bex, Floris and Douglas Walton (2016). ‘Combining explanation and argumentation in dialogue’. In: *Argument & Computation* 7.1, pp. 55–68. DOI: [10.3233/AAC-160001](https://doi.org/10.3233/AAC-160001).
- Black, Elizabeth and Anthony Hunter (2008). ‘Using enthymemes in an inquiry dialogue system’. In: *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pp. 437–444.
- Bocklisch, Tom et al. (2024). *Task-Oriented Dialogue with In-Context Learning*. arXiv: [2402.12234](https://arxiv.org/abs/2402.12234) [cs.CL].
- Booshehri, Meisam, Hendrik Buschmeier and Philipp Cimiano (2024). ‘A Model of Factors Contributing to the Success of Dialogical Explanations’. In: *Proceedings of the 26th International Conference on Multimodal Interaction*. ICMI ’24. San Jose, Costa Rica: Association for Computing Machinery, pp. 373–381. ISBN: 9798400704628. DOI: [10.1145/3678957.3685744](https://doi.org/10.1145/3678957.3685744).
- Bordt, Sebastian et al. (2024). ‘Data Science with LLMs and Interpretable Models’. In: *XAI4Sci Workshop at AAAI-24*.
- Braines, Dave, Alun Preece and Dan Harborne (Nov. 2018). ‘Multimodal explanations for AI-based multisensor fusion’. In: *Proc. NATO SET-262 RSM on Artificial Intelligence for Military Multisensor Fusion Engines*. Budapest, Hungary.
- Bratko, Ivan (1986). *Prolog Programming for Artificial Intelligence*. Reading, Massachusetts: Addison-Wesley.

- Breitholtz, Ellen (2020). *Enthymemes and Topoi in Dialogue: The Use of Common Sense Reasoning in Conversation*. Leiden, The Netherlands: Brill. ISBN: 978-90-04-43679-4. DOI: <https://doi.org/10.1163/9789004436794>. URL: <https://brill.com/view/title/58383>.
- Brown, Tom et al. (2020). ‘Language models are few-shot learners’. In: *Advances in neural information processing systems*. Vol. 33, pp. 1877–1901.
- Buchanan, Bruce G. and Edward H. Shortliffe, eds. (1984). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley. ISBN: 978-0201101720.
- Buhl, Heike M., Josephine B. Fisher and Katharina J. Rohlfing (2024). ‘Changes in Partner Models — Effects of Adaptivity in the Course of Explanations’. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 46. Cognitive Science Society. URL: <https://escholarship.org/uc/item/5bb7p5pt>.
- Burrell, Jenna (2016). ‘How the machine ‘thinks’: Understanding opacity in machine learning algorithms’. In: *Big Data & Society* 3.1. DOI: [10.1177/2053951715622512](https://doi.org/10.1177/2053951715622512).
- Buschmeier, Hendrik et al. (2025). ‘Forms of understanding for XAI-Explanations’. In: *Cognitive Systems Research* 94. ISSN: 1389-0417. DOI: <https://doi.org/10.1016/j.cogsys.2025.101419>. URL: <https://www.sciencedirect.com/science/article/pii/S1389041725000993>.
- Bussemeyer, Jerome R et al. (2013). ‘Learning functional relations based on experience with input–output pairs by humans and artificial neural networks’. In: *Knowledge concepts and categories*. Psychology Press, pp. 405–437.
- Capel, Tara and Margot Brereton (2023). ‘What is human-centered about human-centered AI? A map of the research landscape’. In: *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–23.
- Castagna, Federico, Alexandra Garton et al. (2023). ‘EQRbot: A chatbot delivering EQR argument-based explanations’. In: *Frontiers in Artificial Intelligence* 6. ISSN: 2624-8212. DOI: [10.3389/frai.2023.1045614](https://doi.org/10.3389/frai.2023.1045614). URL: <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1045614>.
- Castagna, Federico, Nadin Kökciyan et al. (2024). ‘Computational argumentation-based chatbots: a survey’. In: *Journal of Artificial Intelligence Research* 80, pp. 1271–1310.
- Cawsey, Alison (1993a). ‘Planning interactive explanations’. In: *International Journal of Man-Machine Studies* 38.2, pp. 169–199.
- (1993b). ‘User modelling in interactive explanations’. In: *User Modeling and User-Adapted Interaction* 3, pp. 221–247.

- Chandrasekaran, B., Michael C. Tanner and John R. Josephson (Jan. 1989). ‘Explaining Control Strategies in Problem Solving’. In: *IEEE Intelligent Systems* 4.01, pp. 9–15, 19–24. ISSN: 1941-1294. DOI: [10.1109/64.21896](https://doi.org/10.1109/64.21896).
- Chen, Hongshen et al. (2017). ‘A survey on dialogue systems: Recent advances and new frontiers’. In: *Acm Sigkdd Explorations Newsletter* 19.2, pp. 25–35.
- Chromik, Michael et al. (2021). ‘I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI’. In: *Proceedings of the 26th International Conference on Intelligent User Interfaces*. IUI ’21. College Station, TX, USA: Association for Computing Machinery, pp. 307–317. ISBN: 9781450380171. DOI: [10.1145/3397481.3450644](https://doi.org/10.1145/3397481.3450644).
- Cinà, Giovanni et al. (2023). ‘Semantic match: Debugging feature attribution methods in XAI for healthcare’. In: *Conference on Health, Inference, and Learning*. PMLR, pp. 182–190.
- Cooper, Robin (2023). *From perception to communication: a theory of types for action and meaning*. Oxford University Press.
- Cover, Thomas and Peter Hart (1967). ‘Nearest neighbor pattern classification’. In: *IEEE transactions on information theory* 13.1, pp. 21–27.
- Dahlbäck, Nils, Arne Jönsson and Lars Ahrenberg (1993). ‘Wizard of Oz studies: why and how’. In: *Proceedings of the 1st international conference on Intelligent user interfaces*, pp. 193–200.
- Dazeley, Richard et al. (2021). ‘Levels of explainable artificial intelligence for human-aligned conversational explanations’. In: *Artificial Intelligence* 299. ISSN: 0004-3702. DOI: <https://doi.org/10.1016/j.artint.2021.103525>. URL: <https://www.sciencedirect.com/science/article/pii/S000437022100076X>.
- den Boeft, Madelon et al. (Feb. 2017). ‘Negotiating explanations: doctor–patient communication with patients with medically unexplained symptoms—a qualitative analysis’. English. In: *Family Practice* 34.1, pp. 107–113. ISSN: 0263-2136. DOI: [10.1093/fampra/cmw113](https://doi.org/10.1093/fampra/cmw113).
- Dignum, Virginia (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Vol. 2156. Springer.
- Dohnány, Sebastian et al. (Mar. 2026). ‘Technological folie à deux: feedback loops between AI chatbots and mental health’. In: *Nature Mental Health* 4.3, pp. 336–345. ISSN: 2731-6076. DOI: [10.1038/s44220-026-00595-8](https://doi.org/10.1038/s44220-026-00595-8).
- Doorn, William P. T. M. van et al. (Jan. 2021). ‘A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis’. In: *PLOS ONE* 16.1, pp. 1–15. DOI: [10.1371/journal.pone.0245157](https://doi.org/10.1371/journal.pone.0245157).
- Ducrot, Oswald (1988). ‘Topoi et formes topiques’. In: *Bulletin d'études de linguistique française* 22.1, pp. 1–14.

- Ehsan, Upol et al. (2021). ‘Operationalizing human-centered perspectives in explainable AI’. In: *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pp. 1–6.
- Eiband, Malin et al. (2019). ‘The Impact of Placebic Explanations on Trust in Intelligent Systems’. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI EA ’19. Glasgow, Scotland Uk: Association for Computing Machinery, pp. 1–6. ISBN: 9781450359719. DOI: [10.1145/3290607.3312787](https://doi.org/10.1145/3290607.3312787).
- Feldhus, Nils, Ajay Madhavan Ravichandran and Sebastian Möller (2022). *Mediators: Conversational Agents Explaining NLP Model Behavior*. arXiv: [2206.06029](https://arxiv.org/abs/2206.06029) [cs.CL].
- Feustel, Isabel et al. (Sept. 2024). ‘Enhancing Model Transparency: A Dialogue System Approach to XAI with Domain Knowledge’. In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Tatsuya Kawahara et al. Kyoto, Japan: Association for Computational Linguistics, pp. 248–258. DOI: [10.18653/v1/2024.sigdial-1.22](https://doi.org/10.18653/v1/2024.sigdial-1.22). URL: <https://aclanthology.org/2024.sigdial-1.22/>.
- Fine Licht, Karl de and Jenny de Fine Licht (2020). ‘Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy’. In: *AI & society* 35.4, pp. 917–926.
- Fine Licht, Karl de and Anna Folland (2025). ‘AI in Public Decision-Making: A Philosophical and Practical Framework for Assessing and Weighing Harm and Benefit’. In: *Public Administration*. Online version before inclusion in issue. DOI: <https://doi.org/10.1111/padm.70029>.
- Finke, Josefine et al. (2022). ‘(De)Coding Social Practice in the Field of XAI: Towards a Co-constructive Framework of Explanations and Understanding Between Lay Users and Algorithmic Systems’. In: *Artificial Intelligence in HCI*. Ed. by Helmut Degen and Stavroula Ntoa. Cham: Springer International Publishing, pp. 149–160. ISBN: 978-3-031-05643-7.
- Fisher, Josephine B et al. (2023). ‘Exploring the Semantic Dialogue Patterns of Explanations—a Case Study of Game Explanations’. In: *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 35–46.
- Fisher, Josephine B. et al. (Dec. 2022). ‘Exploring Monological and Dialogical Phases in Naturally Occurring Explanations’. In: *KI - Künstliche Intelligenz* 36.3, pp. 317–326. ISSN: 1610-1987. DOI: [10.1007/s13218-022-00787-1](https://doi.org/10.1007/s13218-022-00787-1).
- Floridi, Luciano (2021). ‘Establishing the rules for building trustworthy AI’. In: *Ethics, governance, and policies in artificial intelligence*. Springer, pp. 41–45.
- Forrest, James et al. (2018). ‘Towards making NLG a voice for interpretable machine learning’. In: *Proceedings of The 11th International Natural Language Generation Conference*. Association for Computational Linguistics (ACL).

- Friedman, Jerome H (2001). ‘Greedy function approximation: a gradient boosting machine’. In: *Annals of statistics*, pp. 1189–1232.
- Fritzell, Peter, Johan Mesterton and Olle Hagg (Apr. 2022). ‘Prediction of outcome after spinal surgery—using The Dialogue Support based on the Swedish national quality register’. In: *European Spine Journal* 31.4, pp. 889–900. ISSN: 1432-0932. DOI: [10.1007/s00586-021-07065-y](https://doi.org/10.1007/s00586-021-07065-y).
- Ginzburg, Jonathan (1996). ‘Interrogatives: Questions, facts and dialogue’. In: *The handbook of contemporary semantic theory* 5.18, pp. 359–423.
- (2012). *The interactive stance*. New York, USA: Oxford University Press.
- Girard, Jean-Yves (1987). ‘Linear logic’. In: *Theoretical computer science* 50.1, pp. 1–101.
- Google (2025). *Dialogflow*. URL: <https://dialogflow.cloud.google.com> (visited on 19/09/2025).
- Gopnik, Alison et al. (2004). ‘A Theory of Causal Learning in Children: Causal Maps and Bayes Nets.’ In: *Psychological Review* 111.1, pp. 3–32. DOI: [10.1037/0033-295X.111.1.3](https://doi.org/10.1037/0033-295X.111.1.3).
- Götze, Jana and David Schlangen (2023). “‘Why do you say so?’: Dialogical classification explanations in the wild and elicited through classification games’. In: *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*.
- Grice, Herbert Paul (1975). ‘Logic and conversation’. In: *Syntax and semantics* 3, pp. 43–58.
- Grimmelikhuijsen, Stephan (2023). ‘Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making’. In: *Public Administration Review* 83.2, pp. 241–262. DOI: <https://doi.org/10.1111/puar.13483>.
- Guidotti, Riccardo et al. (2021). ‘Principles of Explainable Artificial Intelligence’. In: *Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications*. Ed. by Moamar Sayed-Mouchaweh. Cham: Springer International Publishing, pp. 9–31. ISBN: 978-3-030-76409-8. DOI: [10.1007/978-3-030-76409-8_2](https://doi.org/10.1007/978-3-030-76409-8_2).
- Gulbrandsen, P, A Finset and BF Jensen (2013). *Lege-pasient-korpus fra Ahus*. Tekstlaboratoriet, ILN, UiO. URL: <http://www.hf.uio.no/iln/english/about/organization/text-laboratory/projects/doctor-patient/index.html>.
- Gunning, David et al. (2021). ‘DARPA’s explainable AI (XAI) program: A retrospective’. In: *Applied AI Letters* 2.4, e61. DOI: <https://doi.org/10.1002/ail2.61>.
- Habermas, Jürgen (1984). *The Theory of Communicative Action*. Boston, MA: Beacon Press.

- (1990). *Moral Consciousness and Communicative Action*. Studies in contemporary German social thought. MIT Press. ISBN: 9780262581189.
- Harmon-Jones, Eddie Ed (2019). *Cognitive dissonance: Reexamining a pivotal theory in psychology*. American Psychological Association.
- Harvey, John H., Ann L. Weber and Terry L. Orbuch (1990). *Interpersonal accounts: A social psychological perspective*. Social psychology and society. Cambridge, MA, US: Basil Blackwell, pp. xiii, 184–xiii, 184. ISBN: 0-631-17592-X (Hardcover).
- Hase, Peter and Mohit Bansal (July 2020). ‘Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?’ In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 5540–5552. DOI: [10.18653/v1/2020.acl-main.491](https://doi.org/10.18653/v1/2020.acl-main.491). URL: <https://aclanthology.org/2020.acl-main.491/>.
- Hastie, Trevor and Robert Tibshirani (1986). ‘Generalized Additive Models’. In: *Statistical Science* 1.3, pp. 297–310.
- He, Gaole, Nilay Aishwarya and Ujwal Gadiraju (2025). ‘Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant’. In: *Proceedings of the 30th International Conference on Intelligent User Interfaces*. IUI ’25. Association for Computing Machinery, pp. 907–924. ISBN: 9798400713064. DOI: [10.1145/3708359.3712133](https://doi.org/10.1145/3708359.3712133).
- Heider, Fritz (2013). *The psychology of interpersonal relations*. Psychology Press.
- Hempel, Carl G. and Paul Oppenheim (1948). ‘Studies in the Logic of Explanation’. In: *Philosophy of Science* 15.2, pp. 135–175. DOI: [10.1086/286983](https://doi.org/10.1086/286983).
- Heritage, John and Douglas W Maynard (2006). *Communication in medical care: Interaction between primary care physicians and patients*. Vol. 20. Cambridge, UK: Cambridge University Press.
- Hernandez-Bocanegra, Diana C and Jürgen Ziegler (2021). ‘Conversational review-based explanations for recommender systems: Exploring users’ query behavior’. In: *CUI 2021-3rd Conference on Conversational User Interfaces*, pp. 1–11.
- (Apr. 2023). ‘Explaining Recommendations through Conversations: Dialog Model and the Effects of Interface Type and Degree of Interactivity’. In: *ACM Trans. Interact. Intell. Syst.* 13.2. ISSN: 2160-6455. DOI: [10.1145/3579541](https://doi.org/10.1145/3579541).
- Hilton, Denis J. (1990). ‘Conversational processes and causal explanation.’ In: *Psychological Bulletin* 107.1, pp. 65–81. DOI: [10.1037/0033-2909.107.1.65](https://doi.org/10.1037/0033-2909.107.1.65).
- Hirschberg, Julia Bell (1985). ‘A theory of scalar implicature’. PhD thesis. University of Pennsylvania.
- Hofmann, Hans (1994). *Statlog (German Credit Data)*. UCI Machine Learning Repository. DOI: [10.24432/C5NC77](https://doi.org/10.24432/C5NC77).
- Hosseini, Seyed Ali (2016). ‘Dialogues incorporating enthymemes and modelling of other agents’ beliefs’. PhD thesis. King’s College London.

- Hudeček, Vojtěch and Ondrej Dusek (Sept. 2023). ‘Are Large Language Models All You Need for Task-Oriented Dialogue?’ In: *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Svetlana Stoyanchev et al. Prague, Czechia: Association for Computational Linguistics, pp. 216–228. DOI: [10.18653/v1/2023.sigdial-1.21](https://doi.org/10.18653/v1/2023.sigdial-1.21). URL: <https://aclanthology.org/2023.sigdial-1.21/>.
- Jacovi, Alon and Yoav Goldberg (July 2020). ‘Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?’ In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 4198–4205. DOI: [10.18653/v1/2020.acl-main.386](https://doi.org/10.18653/v1/2020.acl-main.386). URL: <https://aclanthology.org/2020.acl-main.386/>.
- Janosi, Andras et al. (1989). *Heart Disease*. UCI Machine Learning Repository. DOI: [10.24432/C52P4X](https://doi.org/10.24432/C52P4X).
- Johannesen, Leila Julia, Richard I. Cook and David D. Woods (Dec. 1994). *Grounding explanations in evolving, diagnostic situations*. Final Report, RF Project No. 760678/727376. The Ohio State University.
- John, Oliver P. and Sanjay Srivastava (1999). ‘The Big Five Trait taxonomy: History, measurement, and theoretical perspectives’. In: *Handbook of personality: Theory and research, 2nd ed.* Ed. by Lawrence A Pervin and Oliver P John. New York, NY, US: Guilford Press, pp. 102–138. ISBN: 1-57230-483-9 (Hardcover).
- Jönsson, Arne and Nils Dahlbäck (2000). ‘Distilling dialogues — A method using natural dialogue corpora for dialogue systems development’. In: *6th Applied Natural Language Processing Conference, Seattle, Washington, 2000*. Association for Computational Linguistics Stroudsburg, pp. 44–51.
- Kaczmarek-Majer, Katarzyna et al. (2022). ‘PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries’. In: *Information Sciences* 614, pp. 374–399.
- Kahneman, Daniel (2011). *Thinking, fast and slow*. New York, NY, US: Farrar, Straus and Giroux. ISBN: 0-374-27563-7 (Hardcover).
- Kahneman, Daniel, Olivier Sibony and Cass R Sunstein (2021). *Noise: A flaw in human judgment*. Hachette UK.
- Kamar, Ece (2016). ‘Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence’. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. New York, NY, pp. 4070–4073.
- Kato, Masahiro and Masaaki Imaizumi (2023). *CATE Lasso: Conditional Average Treatment Effect Estimation with High-Dimensional Linear Regression*. arXiv: [2310.16819](https://arxiv.org/abs/2310.16819) [econ.EM].
- Kaur, Davinder et al. (2022). ‘Trustworthy artificial intelligence: a review’. In: *ACM computing surveys (CSUR)* 55.2, pp. 1–38.

- Khaneman, Daniel (2003). 'A perspective on judgement and choice'. In: *American Psychologist* 58, pp. 697–720.
- Kim, Jenia, Henry Maathuis and Danielle Sent (2024). 'Human-centered evaluation of explainable AI applications: a systematic review'. In: *Frontiers in Artificial Intelligence* Volume 7 - 2024. ISSN: 2624-8212. DOI: [10.3389/frai.2024.1456486](https://doi.org/10.3389/frai.2024.1456486). URL: <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1456486>.
- Knoks, Aleks and Thomas Raleigh (2022). 'XAI and philosophical work on explanation: A roadmap'. In: *Proceedings of 1st Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming*. Vol. 3319. CEUR-WS.
- Krauss, Daniel A., Gabriel I. Cook and Lukas Klapatch (2018). 'Risk assessment communication difficulties: An empirical examination of the effects of categorical versus probabilistic risk communication in sexually violent predator decisions'. In: *Behavioral Sciences & the Law* 36.5, pp. 532–553. DOI: <https://doi.org/10.1002/bsl.2379>.
- Krishna, Satyapriya et al. (2025). *The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective*. arXiv: [2202.01602](https://arxiv.org/abs/2202.01602) [cs.LG].
- Kuhlen, Anna K and Susan E Brennan (2013). 'Language in dialogue: When confederates might be hazardous to your data'. In: *Psychonomic bulletin & review* 20.1, pp. 54–72.
- Kuźba, Michał and Przemysław Biecek (2020). 'What Would You Ask the Machine Learning Model? Identification of User Needs for Model Explanations Based on Human-Model Conversations'. In: *ECML PKDD 2020 Workshops*. Ed. by Irena Koprinska et al. Cham: Springer International Publishing, pp. 447–459. ISBN: 978-3-030-65965-3.
- Lakkaraju, Himabindu, Stephen H Bach and Jure Leskovec (2016). 'Interpretable decision sets: A joint framework for description and prediction'. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1675–1684.
- Lakkaraju, Himabindu, Dylan Slack et al. (2022). *Rethinking Explainability as a Dialogue: A Practitioner's Perspective*. arXiv: [2202.01875](https://arxiv.org/abs/2202.01875) [cs.LG].
- Lanham, Tamera et al. (2023). *Measuring Faithfulness in Chain-of-Thought Reasoning*. arXiv: [2307.13702](https://arxiv.org/abs/2307.13702) [cs.AI].
- Lappin, Shalom (2021). *Deep Learning and linguistic representation*. Chapman and Hall/CRC.
- Larman, Craig and Victor R Basili (2003). 'Iterative and incremental developments: A brief history'. In: *Computer* 36.6, pp. 47–56.
- Larson, Stefan et al. (2019). *An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction*. arXiv: [1909.02027](https://arxiv.org/abs/1909.02027) [cs.CL].

- Larsson, Staffan (2002). ‘Issue-based dialogue management’. PhD thesis. Gothenburg, Sweden.
- Larsson, Staffan, Alexander Berman et al. (Oct. 2000). *TrindiKit 2.0 Manual*. Deliverable D5.3. Trindi Project.
- Larsson, Staffan, Lena Santamarta and Arne Jönsson (2000). ‘Using the process of distilling dialogues to understand dialogue systems’. In: *Proceedings of 6th International Conference on Spoken Language Processing (IC-SLP2000/INTERSPEECH2000)*, pp. 374–377.
- Larsson, Staffan and David R Traum (2000). ‘Information state and dialogue management in the TRINDI dialogue move engine toolkit’. In: *Natural language engineering* 6.3-4, pp. 323–340.
- LeCun, Yann, Yoshua Bengio and Geoffrey Hinton (May 2015). ‘Deep learning’. In: *Nature* 521.7553, pp. 436–444. ISSN: 1476-4687. DOI: [10.1038/nature14539](https://doi.org/10.1038/nature14539).
- Liao, Q Vera, Daniel Gruen and Sarah Miller (2020). ‘Questioning the AI: Informing Design Practices for Explainable AI User Experiences’. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–15.
- Liao, Q Vera and Kush R. Varshney (2022). *Human-Centered Explainable AI (XAI): From Algorithms to User Experiences*. arXiv: [2110.10790](https://arxiv.org/abs/2110.10790) [cs.AI].
- Liedeker, Felix et al. (2025). *A User Study Evaluating Argumentative Explanations in Diagnostic Decision Support*. arXiv: [2505.10188](https://arxiv.org/abs/2505.10188) [cs.AI].
- Lind, E. Allan, Ruth Kanfer and P. Christopher Earley (1990). ‘Voice, control, and procedural justice: Instrumental and noninstrumental concerns in fairness judgments.’ In: *Journal of Personality and Social Psychology* 59.5, pp. 952–959. DOI: [10.1037/0022-3514.59.5.952](https://doi.org/10.1037/0022-3514.59.5.952).
- Lindgren, Stefan and Knut Aspegren (2004). *Kliniska färdigheter: informationsutbytet mellan patient och läkare*. Studentlitteratur AB.
- Lipton, Zachary C (2018). ‘The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.’ In: *Queue* 16.3, pp. 31–57.
- Lombrozo, Tania (2006). ‘The structure and function of explanations’. In: *Trends in cognitive sciences* 10.10, pp. 464–470.
- Lombrozo, Tania and Nicholas Z Gwynne (Sept. 2014). ‘Explanation and inference: Mechanistic and functional explanations guide property generalization’. In: *Frontiers in human neuroscience* 8.
- Lucas, Christopher G et al. (2015). ‘A rational model of function learning’. In: *Psychonomic bulletin & review* 22.5, pp. 1193–1215.
- Lundberg, Scott M. and Su-In Lee (2017). ‘A unified approach to interpreting model predictions’. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., pp. 4768–4777. ISBN: 9781510860964.

- Madumal, Prashan et al. (2019). ‘A Grounded Interaction Protocol for Explainable Artificial Intelligence’. In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1033–1041.
- Malle, Bertram F (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press.
- Maraev, Vladislav, Alexander Berman and Staffan Larsson (Sept. 2025). ‘Combining Information State Update, Harel Statecharts and LLMs for controllable and flexible Conversational AI’. In: *Proceedings of the 2025 CLASP Conference on Language models And RePresentations (LARP)*. Ed. by Nikolai Ilinykh, Matias Appelgren and Erik Lagerstedt. Gothenburg, Sweden: Association for Computational Linguistics, pp. 28–37. ISBN: 979-8-89176-249-7. URL: <https://aclanthology.org/2025.clasp-main.3/>.
- Maraev, Vladislav, Jean-Philippe Bernardy and Jonathan Ginzburg (2020). ‘Dialogue management with linear logic: the role of metavariables in questions and clarifications’. In: *Traitement Automatique des Langues* 61.3, pp. 43–67.
- Marques-Silva, Joao and Alexey Ignatiev (Apr. 2023). ‘No silver bullet: interpretable ML models must be explained’. In: *Frontiers in artificial intelligence* 6.
- Maruf, Sameen et al. (June 2023). ‘Influence of context on users’ views about explanations for decision-tree predictions’. In: *Computer Speech & Language* 81. ISSN: 0885-2308. DOI: <https://doi.org/10.1016/j.csl.2023.101483>. URL: <https://www.sciencedirect.com/science/article/pii/S0885230823000025>.
- McClure, John and Denis Hilton (1997). ‘For you can’t always get what you want: When preconditions are better explanations than goals’. In: *British Journal of Social Psychology* 36.2, pp. 223–240.
- McCullagh, Peter and J.A. Nelder (1989). *Generalized linear models. Second edition*. Routledge.
- McSherry, David (2001). ‘Interactive case-based reasoning in sequential diagnosis’. In: *Applied Intelligence* 14, pp. 65–76.
- McTear, Michael (Jan. 2020). *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. eng. Vol. 00048. Synthesis Lectures on Human Language Technologies. Cham, Switzerland: Springer. ISBN: 9781636390314. URL: <https://research.ebsco.com/linkprocessor/plink?id=c23dd26b-2580-393b-9edc-480449cef51>.
- Meinke, Alexander et al. (2025). *Frontier Models are Capable of In-context Scheming*. arXiv: 2412.04984 [cs.AI].
- Melchiorre, Alessandro B. and Markus Schedl (2020a). ‘Personality Correlates of Music Audio Preferences for Modelling Music Listeners’. In: *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. New

- York, NY, USA: Association for Computing Machinery, pp. 313–317. ISBN: 9781450368612.
- Melchiorre, Alessandro B. and Markus Schedl (2020b). *Personality Correlates of Music Audio Preferences for Modelling Music Listeners: Public repository*. URL: <https://gitlab.cp.jku.at/alessandro/pers-corr> (visited on 14/11/2025).
- Mercier, Hugo and Dan Sperber (2017). *The enigma of reason*. Harvard University Press.
- Messner, Martin (2009). ‘The limits of accountability’. In: *Accounting, Organizations and Society* 34.8, pp. 918–938.
- Microsoft (2025). *Language Understanding (LUIS)*. URL: <https://www.luis.ai> (visited on 19/09/2025).
- Mikalef, Patrick et al. (2022). ‘Thinking responsibly about responsible AI and ‘the dark side’ of AI’. In: *European Journal of Information Systems* 31.3, pp. 257–268. DOI: [10.1080/0960085X.2022.2026621](https://doi.org/10.1080/0960085X.2022.2026621).
- Miller, Tim (2019). ‘Explanation in artificial intelligence: Insights from the social sciences’. In: *Artificial intelligence* 267, pp. 1–38.
- (2023). ‘Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI’. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’23. Chicago, IL, USA: Association for Computing Machinery, pp. 333–342. ISBN: 9798400701924. DOI: [10.1145/3593013.3594001](https://doi.org/10.1145/3593013.3594001).
- Mindlin, Dmitry, Fabian Beer et al. (2025). ‘Beyond one-shot explanations: a systematic literature review of dialogue-based xAI approaches’. In: *Artificial Intelligence Review* 58.3, p. 81. ISSN: 1573-7462. DOI: [10.1007/s10462-024-11007-7](https://doi.org/10.1007/s10462-024-11007-7).
- Mindlin, Dmitry, Meisam Booshehri and Philipp Cimiano (2026). ‘Towards Co-Constructed Explanations: A Multi-Agent Reasoning-Based Conversational System for Adaptive Explanations’. In: *Proceedings of the 13th International Conference on Human-Agent Interaction*. HAI ’25. Association for Computing Machinery, pp. 148–157. ISBN: 9798400721786. DOI: [10.1145/3765766.3765768](https://doi.org/10.1145/3765766.3765768).
- Mindlin, Dmitry, Amelie Sophie Robrecht et al. (2024). ‘Measuring User Understanding in Dialogue-Based XAI Systems’. In: *27th European Conference on Artificial Intelligence (ECAI 2024), Including PAIS 2024*. Vol. 392. Frontiers in Artificial Intelligence and Applications. Santiago de Compostela, Spain: IOS Press, pp. 1148–1155. ISBN: 978-1-64368-548-9. DOI: [10.3233/FAIA240608](https://doi.org/10.3233/FAIA240608).
- Mittelstadt, Brent (2019). ‘Principles alone cannot guarantee ethical AI’. In: *Nature machine intelligence* 1.11, pp. 501–507.
- Mittelstadt, Brent, Chris Russell and Sandra Wachter (2019). ‘Explaining Explanations in AI’. In: *Proceedings of the Conference on Fairness, Accountability, and*

- Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, pp. 279–288. ISBN: 9781450361255. DOI: [10.1145/3287560.3287574](https://doi.org/10.1145/3287560.3287574).
- Modgil, Sanjay and Henry Prakken (2014). ‘The ASPIC+ framework for structured argumentation: a tutorial’. In: *Argument & Computation* 5.1, pp. 31–62.
- Molnar, Christoph (2025). *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable (3rd ed.)* Munich, Germany: Christoph Molnar. ISBN: 978-3911578035.
- Moore, Johanna D. and Catherine L. Paris (1993). ‘Planning text for advisory dialogues: capturing intentional and rhetorical information’. In: *Computational Linguistics* 19.4, pp. 651–694.
- Mothilal, Ramaravind K., Amit Sharma and Chenhao Tan (2020). ‘Explaining machine learning classifiers through diverse counterfactual explanations’. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, pp. 607–617. ISBN: 9781450369367. DOI: [10.1145/3351095.3372850](https://doi.org/10.1145/3351095.3372850).
- Nguyen, Van Bach, Jörg Schlötterer and Christin Seifert (2023). ‘From Black Boxes to Conversations: Incorporating XAI in a Conversational Agent’. In: *Explainable Artificial Intelligence*. Ed. by Luca Longo. Cham: Springer Nature Switzerland, pp. 71–96. ISBN: 978-3-031-44070-0.
- (2024). ‘XAgent: A Conversational XAI Agent Harnessing the Power of Large Language Models’. In: *xAI-2024: Late-breaking Work, Demos and Doctoral Consortium (Proceedings of the 2nd World Conference on eXplainable Artificial Intelligence)*. Ed. by Luca Longo, Weiru Liu and Grégoire Montavon. Vol. 3793. CEUR Workshop Proceedings. Valtta, Malta: CEUR-WS.org, pp. 273–280. URL: https://ceur-ws.org/Vol-3793/paper_35.pdf.
- Nisbett, Richard E. and Timothy D. Wilson (1977). ‘Telling more than we can know: Verbal reports on mental processes.’ In: *Psychological Review* 84.3, pp. 231–259. DOI: [10.1037/0033-295X.84.3.231](https://doi.org/10.1037/0033-295X.84.3.231).
- Olsen, Henrik Palmer, Jacob Livingston Slosser and Thomas Troels Hildebrandt (Apr. 2020). ‘What’s in the box? The legal requirement to explain computationally aided decision-making in public administration’. In: *Constitutional Challenges in the Algorithmic Society*. OUP, University of Copenhagen Faculty of Law Research Paper No. 2020-97. URL: <https://ssrn.com/abstract=3580128>.
- OpenAI (2025). *ChatGPT*. URL: <https://chatgpt.com/> (visited on 08/10/2025).
- Østergaard, Søren Dinesen (Aug. 2023). ‘Will Generative Artificial Intelligence Chatbots Generate Delusions in Individuals Prone to Psychosis?’ In: *Schizophrenia Bulletin* 49.6, pp. 1418–1419. ISSN: 0586-7614. DOI: [10.1093/schbul/sbad128](https://doi.org/10.1093/schbul/sbad128).
- Pantanowitz, Liron et al. (Mar. 2024). ‘Nongenerative Artificial Intelligence in Medicine: Advancements and Applications in Supervised and Unsuper-

- vised Machine Learning’. In: *Modern Pathology* 38.3. ISSN: 0893-3952. DOI: <https://doi.org/10.1016/j.modpat.2024.100680>. URL: <https://www.sciencedirect.com/science/article/pii/S0893395224002606>.
- Pearl, Judea (2009). *Causality*. Cambridge university press.
- Pedregosa, F. et al. (2011). ‘Scikit-learn: Machine Learning in Python’. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
- Peirce, Charles Sanders (1992). *The essential peirce: Selected philosophical writings (1893-1913)*. Vol. 2. Indiana University Press.
- Peräkylä, Anssi (1997). ‘Conversation analysis: a new model of research in doctor–patient communication’. In: *Journal of the Royal society of Medicine* 90.4, pp. 205–208.
- (2011). ‘Validity in Research on Naturally Occurring Social Interaction’. In: *Qualitative Research: Issues of Theory, Method and Practice. Third Edition*. Ed. by David Silverman. London: Sage, pp. 365–382.
- Phillips, P Jonathon et al. (2020). *Four principles of explainable artificial*. Tech. rep. NIST interagency report.
- Polanyi, Michael (1966). *The Tacit Dimension*. Garden City, NY: Doubleday.
- Prakken, Henry (Dec. 2005). ‘Coherence and Flexibility in Dialogue Games for Argumentation’. In: *Journal of Logic and Computation* 15.6, pp. 1009–1040. ISSN: 0955-792X. DOI: [10.1093/logcom/exi046](https://doi.org/10.1093/logcom/exi046).
- Pya, Natalya and Simon N Wood (2015). ‘Shape constrained additive models’. In: *Statistics and computing* 25, pp. 543–559.
- Rai, Daking et al. (2025). *A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models*. arXiv: [2407.02646 \[cs.AI\]](https://arxiv.org/abs/2407.02646).
- Reiter, Ehud and Robert Dale (2000). *Building Natural Language Generation Systems*. Cambridge, UK: Cambridge university press.
- Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin (2016). ‘“Why Should I Trust You?”: Explaining the Predictions of Any Classifier’. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.
- Rohlfing, Katharina J. et al. (2021). ‘Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems’. In: *IEEE Transactions on Cognitive and Developmental Systems* 13.3, pp. 717–728. DOI: [10.1109/TCDS.2020.3044366](https://doi.org/10.1109/TCDS.2020.3044366).
- Rozenblit, Leonid and Frank Keil (2002). ‘The misunderstood limits of folk science: An illusion of explanatory depth’. In: *Cognitive science* 26.5, pp. 521–562.
- Rudin, Cynthia (2019). ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’. In: *Nature Machine Intelligence* 1.5, pp. 206–215.

- Rudin, Cynthia et al. (2022). ‘Interpretable machine learning: Fundamental principles and 10 grand challenges’. In: *Statistics Surveys* 16, pp. 1–85. DOI: [10.1214/21-SS133](https://doi.org/10.1214/21-SS133).
- Russell, Stuart (2019). *Human compatible: AI and the problem of control*. New York, NY, USA: Viking.
- Russell, Stuart and Peter Norvig (2022). *Artificial Intelligence: A Modern Approach*. 4th. Pearson. URL: <https://aima.cs.berkeley.edu/>.
- Sacks, Harvey, Emanuel A. Schegloff and Gail Jefferson (1974). ‘A Simplest Systematics for the Organization of Turn-Taking for Conversation’. In: *Language* 50.4, pp. 696–735.
- Sanders, Elizabeth B-N and Pieter Jan Stappers (2008). ‘Co-creation and the new landscapes of design’. In: *Co-design* 4.1, pp. 5–18.
- Sanna, Lawrence J. and Norbert Schwarz (2004). ‘Integrating Temporal Biases: The Interplay of Focal Thoughts and Accessibility Experiences’. In: *Psychological Science* 15.7, pp. 474–481. DOI: [10.1111/j.0956-7976.2004.00704.x](https://doi.org/10.1111/j.0956-7976.2004.00704.x).
- Sassoon, Isabel et al. (2019). ‘Explainable Argumentation for Wellness Consultation’. In: *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Ed. by Davide Calvaresi et al. Cham: Springer International Publishing, pp. 186–202. ISBN: 978-3-030-30391-4.
- Schindler, Carolin et al. (May 2025). ‘Automatic Generation of Structured Domain Knowledge for Dialogue-based XAI Systems’. In: *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*. Ed. by Maria Ines Torres et al. Bilbao, Spain: Association for Computational Linguistics, pp. 1–11. ISBN: 979-8-89176-248-0. URL: <https://aclanthology.org/2025.iwds-1.1/>.
- Schneiderhan, Erik and Shamus Khan (2008). ‘Reasons and Inclusion: The Foundation of Deliberation’. In: *Sociological Theory* 26.1, pp. 1–24. DOI: [10.1111/j.1467-9558.2008.00316.x](https://doi.org/10.1111/j.1467-9558.2008.00316.x).
- Scurich, Nicholas (2018). ‘The case against categorical risk estimates’. In: *Behavioral Sciences & the Law* 36.5, pp. 554–564. DOI: <https://doi.org/10.1002/bsl.2382>.
- Searle, John R (1969). *Speech acts: An essay in the philosophy of language*. Cambridge, UK: Cambridge university press.
- Seedhouse, Paul (2005). ‘Conversation analysis as research methodology’. In: *Applying conversation analysis*. London: Palgrave Macmillan, pp. 251–266.
- Shaheen, Qurat-ul-ain, Alice Toniolo and Juliana K. F. Bowles (2020). ‘Dialogue Games for Explaining Medication Choices’. In: *Rules and Reasoning: 4th International Joint Conference*. Ed. by Víctor Gutiérrez-Basulto et al. Cham: Springer International Publishing, pp. 97–111. ISBN: 978-3-030-57977-7.
- Shneiderman, Ben (2022). *Human-centered AI*. Oxford University Press.
- Shortliffe, Edward H. (1976). *Computer-based medical consultations: MYCIN*. Elsevier.

- Siau, Keng and Weiyu Wang (2020). ‘Artificial intelligence (AI) ethics: ethics of AI and ethical AI’. In: *Journal of Database Management (JDM)* 31.2, pp. 74–87.
- Simonsohn, Uri (2007). ‘Clouds make nerds look good: field evidence of the impact of incidental factors on decision making’. In: *Journal of Behavioral Decision Making* 20.2, pp. 143–152. DOI: <https://doi.org/10.1002/bdm.545>.
- Sklar, Elizabeth I and Mohammad Q Azhar (2018). ‘Explanation through argumentation’. In: *Proceedings of the 6th International Conference on Human-Agent Interaction*, pp. 277–285.
- Slack, Dylan et al. (2023). ‘Explaining machine learning models with interactive natural language conversations using TalkToModel’. In: *Nature Machine Intelligence* 5.8, pp. 873–883.
- Slugoski, Ben R et al. (1993). ‘Attribution in conversational context: Effect of mutual knowledge on explanation-giving’. In: *European Journal of Social Psychology* 23.3, pp. 219–238.
- Sokol, Kacper and Peter Flach (July 2018). ‘Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant’. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, pp. 5868–5870. DOI: [10.24963/ijcai.2018/865](https://doi.org/10.24963/ijcai.2018/865).
- (June 2020). ‘One Explanation Does Not Fit All’. In: *KI - Künstliche Intelligenz* 34.2, pp. 235–250. ISSN: 1610-1987. DOI: [10.1007/s13218-020-00637-y](https://doi.org/10.1007/s13218-020-00637-y).
- (Sept. 2024). ‘Interpretable representations in explainable AI: from theory to practice’. In: *Data Mining and Knowledge Discovery* 38.5, pp. 3102–3140. ISSN: 1573-756X. DOI: [10.1007/s10618-024-01010-5](https://doi.org/10.1007/s10618-024-01010-5).
- Spotify (2022). *Spotify Web API*. URL: <https://developer.spotify.com/documentation/web-api/reference> (visited on 30/05/2022).
- Stalnaker, Robert C. (1972). ‘Pragmatics’. In: *Semantics of Natural Language*. Ed. by Donald Davidson and Gilbert Harman. Dordrecht: Springer Netherlands, pp. 380–397. ISBN: 978-94-010-2557-7. DOI: [10.1007/978-94-010-2557-7_11](https://doi.org/10.1007/978-94-010-2557-7_11).
- Stepin, Ilia et al. (2020). ‘Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers’. In: *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8. DOI: [10.1109/FUZZ48607.2020.9177629](https://doi.org/10.1109/FUZZ48607.2020.9177629).
- Strauss, Anselm and Juliet Corbin (1990). *Basics of qualitative research*. Thousand Oaks, CA: Sage publications.
- Svenska Dagbladet (14th Aug. 2025). *Ebba Busch använde falskt AI-citat – ber om ursäkt*. URL: <https://www.svd.se/a/RzWB8A/ebba-busch-anvande-falskt-ai-citat-ber-om-ursakt> (visited on 14/11/2025).

- Szegedy, Christian et al. (2014). *Intriguing properties of neural networks*. arXiv: 1312.6199 [cs.CV].
- Tedeschi, James T and Marc Reiss (1981). ‘Verbal strategies in impression management’. In: *The psychology of ordinary explanations of social behavior* 271.309, pp. 792–799.
- Todorov, Alexander, Mansur Lalljee and William Hirst (2000). ‘Communication context, explanation, and social judgment’. In: *European Journal of Social Psychology* 30.2, pp. 199–209.
- Toulmin, Stephen E (2003). *The uses of argument*. Cambridge, UK: Cambridge university press.
- Turnbull, William (1986). ‘Everyday explanation: The pragmatics of puzzle resolution’. In: *Journal for the Theory of Social Behaviour* 16.2, pp. 141–160.
- Turpin, Miles et al. (2023). ‘Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting’. In: vol. 36, pp. 74952–74965.
- Tversky, Amos and Daniel Kahneman (1974). ‘Judgment under Uncertainty: Heuristics and Biases’. In: *Science* 185.4157, pp. 1124–1131. DOI: [10.1126/science.185.4157.1124](https://doi.org/10.1126/science.185.4157.1124).
- Tyler, Tom R (2006). *Why people obey the law*. Princeton university press.
- U.S. District Court for the Southern District of New York (2023). *Mata v. Avianca, Inc.* URL: <https://www.courtlistener.com/opinion/9885417/mata-v-avianca-inc/> (visited on 03/10/2025).
- Uceda-Sosa, Rosario et al. (2024). *Reasoning about concepts with LLMs: Inconsistencies abound*. arXiv: 2405.20163 [cs.CL].
- Van Eemeren, Frans H, Bart Garsen and Nanon Labrie (2021). *Argumentation between doctors and patients: Understanding clinical argumentative discourse*. John Benjamins Publishing Company.
- Vasconcelos, Helena et al. (2022). ‘When do XAI methods work? A cost-benefit approach to human-AI collaboration’. In: *CHI Workshop on Trust and Reliance in AI-Human Teams*, pp. 1–15.
- Vaswani, Ashish et al. (2017). ‘Attention is all you need’. In: *Advances in Neural Information Processing Systems*. Vol. 30, pp. 5999–6009.
- Wahde, Mattias and Marco Virgolin (2023). ‘DAISY: An Implementation of Five Core Principles for Transparent and Accountable Conversational AI’. In: *International Journal of Human–Computer Interaction* 39.9, pp. 1856–1873. DOI: [10.1080/10447318.2022.2081762](https://doi.org/10.1080/10447318.2022.2081762).
- Waldmann, Michael R. (1996). ‘Knowledge-Based Causal Induction’. In: *Causal Learning*. Ed. by David R. Shanks, Douglas L. Medin and Keith J. Holyoak. Vol. 34. Psychology of Learning and Motivation. Academic Press, pp. 47–88.

- DOI: [https://doi.org/10.1016/S0079-7421\(08\)60558-7](https://doi.org/10.1016/S0079-7421(08)60558-7). URL: <https://www.sciencedirect.com/science/article/pii/S0079742108605587>.
- Walker, Marilyn A (1996). ‘The effect of resource limits and task complexity on collaborative planning in dialogue’. In: *Artificial Intelligence* 85.1-2, pp. 181–243.
- Wang, Danding et al. (2019). ‘Designing theory-driven user-centric explainable AI’. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–15.
- Wason, Peter C and J St BT Evans (1974). ‘Dual processes in reasoning?’ In: *Cognition* 3.2, pp. 141–154.
- Wei, Jason et al. (2022). ‘Chain-of-thought prompting elicits reasoning in large language models’. In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 24824–24837.
- Weiner, JL (1980). ‘BLAH, a system which explains its reasoning’. In: *Artificial intelligence* 15.1-2, pp. 19–48.
- Weiser, Benjamin (May 2023). *Here’s what happens when your lawyer uses ChatGPT*. URL: <https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html> (visited on 10/10/2025).
- Wielemaker, Jan et al. (2012). ‘SWI-Prolog’. In: *Theory and Practice of Logic Programming* 12.1-2, pp. 67–96. ISSN: 1471-0684.
- Wijekoon, Anjana et al. (2024). *Tell me more: Intent Fulfilment Framework for Enhancing User Experiences in Conversational XAI*. arXiv: 2405.10446 [cs.HC].
- Winograd, Terry (1971). ‘Procedures as a representation for data in a computer program for understanding natural language’. PhD thesis. Massachusetts Institute of Technology.
- Xydis, Andreas et al. (2020). ‘Enthymemes in dialogues’. In: *Computational Models of Argument*. IOS Press, pp. 395–402.
- YAML Language Development Team (2021). *YAML Ain’t Markup Language (YAML™) Version 1.2 Specification (Revision 1.2.2)*. URL: <https://yaml.org/spec/1.2/spec.html> (visited on 22/09/2025).
- Yudkowsky, Eliezer and Nate Soares (2025). *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Little, Brown and Company. ISBN: 978-0-316-59564-3.
- Zhang, Wan and Jing Zhang (2025). ‘Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review’. In: *Mathematics* 13.5. ISSN: 2227-7390. DOI: [10.3390/math13050856](https://doi.org/10.3390/math13050856). URL: <https://www.mdpi.com/2227-7390/13/5/856>.